AN ARTICULATORY-ACOUSTIC INVESTIGATION OF TIMING AND COORDINATION IN THE FLUENT SPEECH OF PEOPLE WHO STAMMER

CORNELIA J HEYDE

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

CLINICAL AUDIOLOGY, SPEECH AND LANGUAGE RESEARCH CENTRE QUEEN MARGARET UNIVERSITY

2019

DECLARATION

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Cornelia J Heyde

ABSTRACT AND LAY SUMMARY

This thesis investigates Wingate's Fault-Line hypothesis (1988) which suggests that disfluencies in people who stammer (PWS) result from a deficit in transition from consonant to vowel (CV) thereby implying that stammering as a motor-control disorder would affect transitions even when not perceptually salient. To test this proposal, we explored the perceptually fluent speech of PWS using instrumental analysis (ultrasound and acoustic) to determine the underlying pervasiveness of disfluencies in this group as compared to people who do not stammer (PNS).

Following fluency screening of recorded utterances, we applied acoustic and articulatory analysis techniques to perceptually fluent utterances of 9 PWS and 9 typical speakers in order to identify indicators of disfluency in the transition from syllable onsets to the following vowel. Measures of acoustic duration, locus equation and formant slope offer insights into timing and degree of coarticulation. The articulatory ultrasound tongue imaging technique moreover provides kinematic information of the tongue. A novel technique was applied to dynamically analyse and quantify the tongue kinematics in transition. This allowed us to treat the perceptually fluent speech of PWS as an ongoing time-situated process.

Both acoustic and articulatory findings indicate by-group differences in timing, whereby PWS are overall slower and more variable in the execution of CV transitions when compared to typical speakers (PNS). The findings from both instrumental approaches also indicate differences in coordination, suggesting that PWS coarticulate to a lesser extent than PNS. Overall, these findings suggest that PWS exhibit a global deficit in CV transition that can be observed in perceptually fluent as well as stammered speech. This is in keeping with the predictions of Wingate's Fault-Line hypothesis.

The fact that the conclusions from the acoustic and articulatory measures are coherent, shows that acoustic measures may be sufficient to act as a proxy for articulatory measures.

ACKNOWLEDGMENTS

There are many people who have played a part in getting me to this point. Foremost, I would like to thank Jim Scobbie for the continuous support of my PhD study and related projects, for his patience, insight and encouragement. I would also like to thank Robin Lickley for posing the hard questions which incentives me to widen my research from various perspectives. Their guidance helped me through the time of research and the writing process of this thesis. I moreover thank Alan Wrench and Steve Cowen for sharing their knowledge on ultrasound recording, as well as Pat Strycharczuk and Ian Finlayson for their statistical input and discussions around it. Thank you to my colleagues at Queen Margaret University who have made this a great experience.

Not to mention the moral support of friends throughout the journey, the value of which should not be underestimated. A huge thank you goes to Elli Drake who has become a great friend and teacher on many levels and to the many friends who patiently listened and helped me through the ups and downs. Thank you, Dermot Fitzsimons, Heather Stevenson, Kat Macapagal, Michelle Foster, Helene Killmer, Line Huck, Laura Bos, Thea Welle, Jette Mögel and last, but not least, my family and in particular my mother Brigitte Heyde and my grandma Ruth Heyde for the many care packages they sent throughout the years and my sister Jana Walzog for always looking after me. You all have been very patient with me. Thank you all for your friendship and support.

Finally, I would like to acknowledge several funding sources that have supported me throughout my PhD. The research presented in this thesis was funded by a Queen Margaret University Research Degree Bursary. This bursary in addition to awards from Queen Margaret University's Princess Alice and Vice Chancellor's Fund, as well as awards from Santander and the AFCP helped me fund my attendance at summer

schools (SPP 2013 and LOT 2013) as well as several conferences (ISSP 2014, ICPhS 2015, Ultrafest 2015, BAAP 2016 and P&P 2016). With the support of LABEX EFL I was invited as a visiting researcher to the LPP at the Sorbonne Nouvelle in 2016, which resulted in very fruitful and enjoyable discussions for this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 Introduction

Stammering is often considered a categorical phenomenon that is either present in speech or not. While assessment of stammering typically relies on perceptual[1] measures, research in the field of stammering has moved toward articulatory methods in recent years.

In this thesis, we explore the articulation of the fluent-sounding speech of people who stammer (PWS). This allows us to investigate whether signs of disfluency are present in even fluent-sounding speech, and whether the fluent speech of PWS differs from that of control speakers (PNS) in this respect. Evidence of articulatory differences between the fluent sounding speech of people who stammer and that of control speakers, would confirm that perceptual salience captures only parts of the signal to be measured, and that the articulatory study of fluent sounding speech may be necessary in order to better understand the underlying nature of stammering. Wingate (1969b, 1988) proposed that the perceptual features of stammering arise due to people who stammer having an underlying difficulty transitioning between syllable onset and the following vowel. The use of articulatory measures in the current study allows us to explore whether there is evidence that people who stammer experience a global difficulty with the CV transition process, which might account for local instances of perceptually identifiable stammering.

[1] In this thesis we will use 'perceptual' availability and 'perceptual' salience to refer to auditory observations that are made without instrumentation.

1.1 Thesis Structure

The thesis is divided into five parts. Chapter 1 presents general concepts about fluency in typical and stammered speech, reviews relevant aspects of current models of speech planning and speech production, and introduces evidence from speech research, which suggests the importance of using articulatory measures when studying speech fluency. There is a specific focus on Wingate's Fault-Line hypothesis (1988) which suggests that people who stammer have difficulty transitioning from syllable onsets to the following vowel (i.e., the CV transition). Implicit to this hypothesis is that the transition difficulty may be present in even the fluent-sounding speech of people who stammer (PWS). This hypothesis informs the rationale of our study in which we employed ultrasound tongue imaging and acoustic analysis to investigate the tongue movements in consonant-vowel (CV) transitions in the fluent speech of PWS, which we compare to those of control speakers (PNS).

The Methods section in Chapter 2 provides details on participant selection, stimuli, and the instrumentation and procedure used to capture data. This section further includes details on screening of the acoustic and articulatory data. Data screening was done to ensure that a) speakers differentiated between the three vowels employed in the study, particularly the low vowel /ɑ/ and the neutral vowel /ə/ and that b) only fluent data would be included in the subsequent acoustic and articulatory analysis.

Chapter 3 presents acoustic analyses of the type traditionally employed for the investigation of coordination patterns in CV transition (here: segment durations, formant slopes and locus equations). The chapter describes the methodology used for the analysis of the acoustic data (where C = /k, t, s, p/ + V = /ɑ, i, ə /) including the segmentation approach applied and measures obtained. It goes on to report and discuss the results in relation to previous research. The findings presented in this chapter provide evidence to suggest that even the fluent speech of PWS differs

from that of PNS. PWS perform with overall longer and more variable acoustic durations. While differences for duration are found overall, differences in variability appear to be driven by larger variability for PWS in closure duration. Coarticulatory measures reveal a group effect for mean locus equations where PWS present with flatter slopes indicating less coarticulation when compared to PNS. Further, measures of variability return significant differences for Locus Equations on lingual consonants as well as for formant slope measures.

Chapter 4 presents a novel approach to articulatory analyses of the data. These include a novel approach to the analysis of tongue kinematics in CV transitions using ultrasound tongue imaging. The methodology is followed by presentation of the results of the articulatory data analysis (where C = /k/): CV transition durations, peak velocity and average speed are presented. Findings support evidence that PWS have difficulty in transitioning between consonant and vowel. Findings reveal no group difference for the movement into the initial consonant (onset stroke). For the movement transitioning from consonantal closure to the subsequent vowel (offset stroke), PWS perform with longer and more variable kinematic movement duration at lower peak velocity.

The last chapter, chapter 5, presents a summary of important findings from the acoustic (chapter 3) and articulatory analyses (chapter 4) and discusses them in relation to previous literature presented in chapter 1. Chapter 5 further presents a description of overt disfluencies and discusses them in relation to the previous analysis (Chapter 3 and Chapter 4) and Wingate's Fault-Line hypothesis. Limitations and suggestions for future research conclude this thesis.

1.2 General Concepts

When researching stammering, there are two questions one cannot avoid asking. There are two elephants in the room: First, what is fluency? How is it different from disfluency? Second, what is the best way to observe and assess fluency / disfluency?

In the following section, we will briefly outline the approach to these questions adopted in the current study.

1.2.1 What is Fluency?

Stammering is a fairly common speech pathology, being experienced by approximately 1 in 20 people at some point during their life (Yairi & Ambrose, 2013). Even though it might be hard to define stammering, most of us will have a reliable idea of what it is at a perceptual level. Stammering affects motor control in those people who experience it. As a primary symptom, it affects the person's speech. In adults who stammer (AWS) physical concomitants were shown to increase with severity (Archibald & De Nil, 1999).

Most of us have probably asked ourselves "Isn't everyone disfluent sometimes?" The answer to the question is probably yes. Fluent speech is at times naturally disfluent with pauses and repetitions interrupting the flow of speech (Brown & McNeill, 1966; Corley, MacGregor, & Donaldson, 2007; Johnson, 1961; R. J. Lickley, 2015; Maclay & Osgood, 1959). Highly natural are silences or hesitations that occur as part of a breathing rhythm (Fuchs, Petrone, Krivokapić, & Hoole, 2013; Rochet-Capellan & Fuchs, 2013). Other hesitations or pauses allow the speaker to organise his upcoming speech (Watanabe, Hirose, Den, & Minematsu, 2008) and repetitions of sounds or syllables may be forms of the speaker restarting to repair what he was about to say (Levelt, 1983; Postma, 2000; Schegloff, Jefferson, & Sacks, 1977).

Stammering, in contrast, is characterised through stammer-like disfluencies that affect the fluency of speech. These stammer-like disfluencies are usually characterised as consisting of three separate phenomena: repetitions ([k\k\kɒfɪ]), prolongations ([ʃ:::u:]), and blocks ([t---ɒp]).

While the typical disfluencies and stammer-like disfluencies are not the same, their surface forms may be similar with overlapping characteristics. So, how are the interruptions in the flow of typical speech different to stammer-like disfluencies in

stammered speech? How can we define disfluencies and eventually stammering robustly?

Various definitions of stammering have been proposed (Brutten & Shoemaker, 1967; Guitar, 2013; Johnson, 1959; Manning & DiLollo, 2017; D. A. Shapiro, 1999; van Riper, 1982; Wingate, 1964, 1984a) – all targeting at the perceptually overt stammer-like disfluencies that affect the fluency of the speaker's speech. Wingate (1984a) distinguishes between what he terms 'typical disfluency' and 'pathological dysfluency' (with "i" and "y" spellings to signal the difference[2]). Typical disfluency he says is the occasional disruption (such as hesitation, interjection, part-word repetition) that occurs in the fluent speech of typical speakers. Pathological disfluencies in contrast are described as "abnormal disruptions" that can only occur as part of pathological speech. The distinction between typical disfluency and pathological disfluency to Wingate is also a perceptual one, where speech appears to be realised either smoothly, with ease and expressively or not (Wingate, 1984a). This is in line with definitions where voluntary disfluencies, such as hesitations in typical speech, often function as pragmatic markers (Norrick, 2009). When placed at turn boundaries and prosodic boundaries they can support the speaker in claiming a turn or maintaining the conversational floor while also allowing for relatively smooth speech. Stammer-like disfluencies in contrast are not controlled to the same extent as they are involuntary and may not be aligned with higher level conversational boundaries (i.e. turn-taking, prosody) causing breakdowns to be more salient and perceived as abnormal disruptions. In addition to their location / linguistic alignment, it can also be said that their inherent characteristics are more pronounced when compared to those of typical disfluencies: First, stammer-like repetitions count more repetitions than mere restarts and second, stammer-like prolongations may exceed by far the duration of what is considered acceptable and

[2] In this thesis we will use 'disfluency' to refer to stammer-like disfluencies, unless otherwise specified.

typical. Third, blocks are not usually observed in typical speech. Overall, disfluencies occur at higher frequency in stammered speech when compared to typical speech (Shriberg, 1995). Once interruptions in the flow of speech are identified as pathological, their frequency and measures of duration are employed to help quantify the severity of the stammer (Boey, Wuyts, Heyning, Bodt, & Heylen, 2007; R. Lickley, 2017; A. I. Shapiro & Decicco, 1982).

The ways stammering manifests itself as well as the group of people who stammer are widely heterogeneous, which poses a problem for identifying and categorising stammered versus fluent speech. This is also reflected in the rather vague ways stammering has been categorised: Pascoe et al. (Pascoe, Stackhouse, & Wells, 2006) categorise stammering using the term speech difficulty, which describes difficulties to produce speech more generally, including difficulty on various linguistic levels (from sound to sentence level) and extending as far as the communication level. Speech sound disorders is a different term that is more internationally established to refer to difficulty in perception as well as production of speech (International Expert Panel on Multilingual Children's Speech, 2012). Following the latter categorisation, speech sound disorders can affect both speech planning and production, which we will briefly discuss in later sections (see 1.3.1 and 1.3.2).

In trying to understand what the underlying cause to these pathological disfluencies could be, stammering research covers a diversity of theoretical fields. The many different theories about causes of stammering range from psychological, genetic, to neurological and linguistic theories (Ambrose, 2004; Ben & Busan, 2014; Büchel & Sommer, 2004; Prasse & Kikano, 2008; Ward, 2018). What they all have in common is the search for an explanation of the disfluencies that affect the speech of people who stammer (as opposed to disfluencies in typical speech). Within the field of stammering research, however, it remains unclear what disfluencies really are and what makes them pathological.

In this thesis we will investigate Wingate's Fault-Line hypothesis in which he suggests that underlying all stammer-like disfluencies is a common mechanism,

which is a phonetic transition deficit (Wingate, 1969b). Wingate refers to as a 'phonetic' transition deficit may be better understood as 'articulatory' transition deficit. What Wingate claims that common to these stammer-like disfluencies is the successful achievement of targets, but the errorful transition between them. While Wingate's hypothesis is conceptually clear, it is lacking in evidence, an issue addressed in the work described in this thesis.

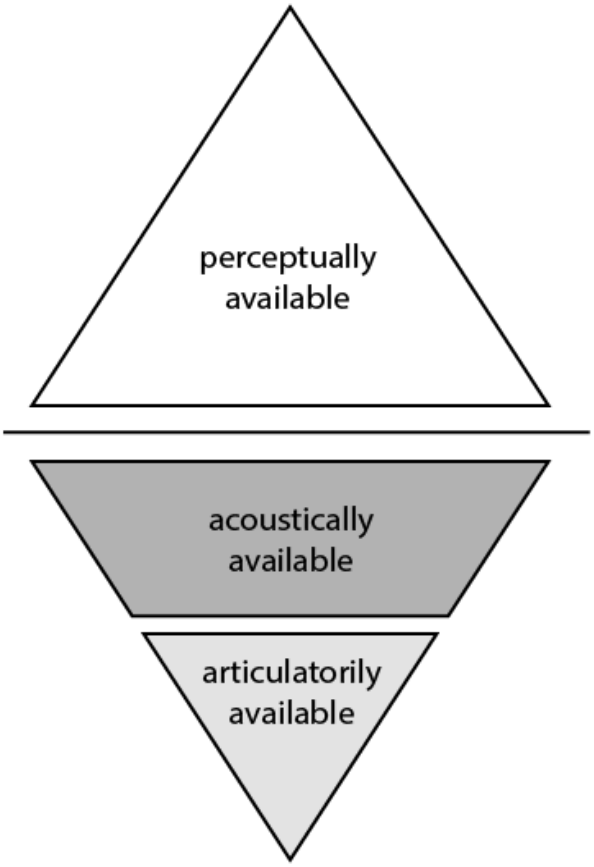


Figure 1 *Schematic representation of overt (top) and covert (bottom) disfluencies and the data required for investigation*

Disfluencies are typically assumed to be local events that intercept the otherwise fluent speech. Following Wingate's assumption of a phonetic transition deficit in people who stammer, the question arises: Are stammered disfluencies indeed local instances of fluency breakdowns surrounded by otherwise fluent speech, as we

perceive them acoustically? Further: Are fluent and disfluent speech indeed easily discernible categories? Or is it rather the case that stammering is a gradient phenomenon that affects speech more globally, where speech does not fall into two easily discernible categories of fluent and disfluent speech, but on a scale from fluent to disfluent speech – a question that has already been raised in the early 1980s (e.g., M. R. Adams & Runyan, 1981).

Thinking of a clear-cut distinction between fluent and disfluent speech, PWS produce potentially very large amounts of perceptually fluent speech. It could also be the case that the speech of people who stammer is affected in a more global manner, but with different portions exhibiting differing degrees of deviance from the fluent speech of typical speakers. The larger the deviance from the typical speech, the greater the chance that it would be recognised at a perceptual level and categorised as pathological (see Figure 1). The inverse might mean that – though to a lesser extent, even perceptually fluent speech carries characteristics of disfluency. In order to better understand stammering it is therefore necessary to obtain a clearer picture of how the perceptually fluent speech of people who stammer compares to that of people who do not stammer.

Therefore, in the current study we will investigate the perceptually fluent speech of people who stammer which we will compare to the speech of control speakers

1.2.2 How is Fluency Different from Disfluency?

It is important to choose carefully the source of information. This section will briefly explore the importance of source for the case of stammering.

For stammering, as for motor control impairments in general, articulatory information is the most direct source to obtain information about motor control mechanisms involved in speech production (Cleland, Scobbie, Roxburgh, & Heyde, 2015; Heyde, Cleland, Scobbie, & Roxburgh, 2017; Wood, Wishart, Hardcastle, Cleland, & Timmins, 2009). The exploration of motor articulation itself is, however, often neglected even though it underlies all acoustic speech output: the acoustic

signal can be regarded to be the result of articulatory effort with air being pushed through the vocal tract where the airstream is modified. Only accurately targeted perturbation of the airflow through configuration of the organs of the vocal tract can lead to the intended acoustic signal. Labial, laryngeal and lingual articulators need to be well orchestrated. Focusing on lingual articulation, phones can be assigned a certain tongue configuration (i.e., gesture). In running speech, these lingual gestures are smoothly connected. In cases where breakdowns can be perceived acoustically, these can also be observed on the articulatory level (Fant, 1970; Lin & Mielke, 2008; Tasko & Greilick, 2010).

There is no one-to-one relationship between information that can extracted from acoustic data and information that can be obtained from articulatory data (Qin & Carreira-Perpiñán, 2007). Often researchers treat acoustic and articulatory information as equally valid. Wingate (Wingate, 1964), for example, claims that from speech perception one can infer the underlying articulatory gesture. This understanding is still popular today, with the consequence that speech motor control impairments such as stammering are typically defined via their acoustically salient / perceptual properties. The translation of acoustic segments into articulatory gestures, however, is not a straightforward as had been implicitly assumed (Ananthakrishnan & Engwall, 2011; Badin, Tarabalka, Elisei, & Bailly, 2010; Kent, 2015). Pape et al. found that “information related to the dynamics of the vocal tract articulators is used by listeners to recover the intended but not uttered vowel target” (Pape, Perrier, Fuchs, & Kandel, 2011, p. 3). This suggests that acoustic judgement is likely to fail in situations where the same targets are achieved, but articulatory trajectories to achieve these targets differ. This finding is particularly important for the study of stammered speech as it shows the limitations of acoustics when trying to explore the underlying articulatory mechanisms.

Perceptual categorisation may be sufficient to characterise a speaker’s output as a whole as including stammering or not. However, in order to identify the underlying

disturbance which can manifest as perceptually identifiable stammering, a more fine-grained analysis may be necessary.

Acoustically salient characteristics are a subset of those observed in articulatory data (see Figure 2). It is important to note here, that 'perceptually salient' and 'acoustically available' are not synonymous. The latter offers details that might be instrumentally measurable, but not necessarily perceivable to the human ear. Both, however, do miss motor information. Acoustic data on its own, therefore, does not capture the whole picture and as such cannot be sufficient when trying to understand the underlying mechanisms of stammering.

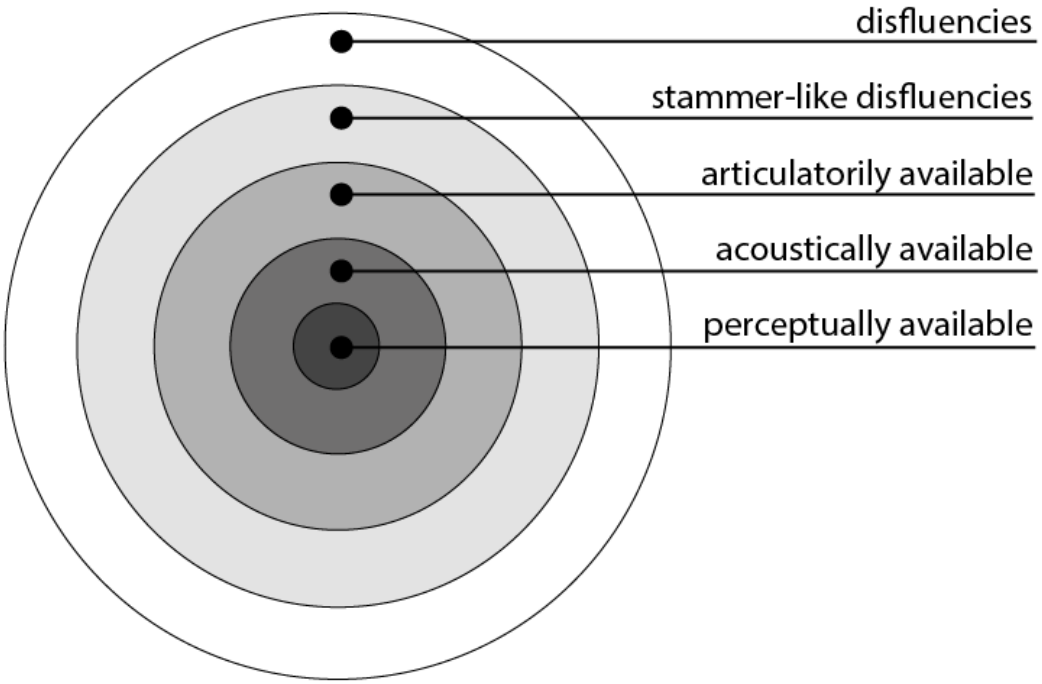


Figure 2 *Schematic representation of disfluencies and how they can be observed*

While all acoustic events can, given the appropriate equipment, be observed articulatorily, the inverse is not necessarily true. Articulatory movements are not necessarily acoustically salient but can be silent and as such they can remain covert to the listener. The same holds for disfluencies. While disfluencies that are

acoustically salient are, by definition, present on the articulatory level (the underlying source of speech), not every instance of disfluency that affects articulation necessarily surfaces onto the acoustic level.

Disfluencies can remain covert if, for example, they are very minor, and the speaker manages to conceal them before they are fully articulated. As such, perceptually fluent speech may contain disfluent features that remain unnoticed by the interlocutor / researcher and can only be observed on the articulatory level (A. Smith, Kelly, Curlee, & Siegel, 1997).

Based on articulatory data a more detailed picture can be obtained providing both overt acoustically salient as well as covert articulatory disfluencies. Articulatory data should therefore be considered as a source of information offering a more holistic view onto the characteristics of speech, which is beneficial to obtain insights into the nature of disfluencies to define and possibly discern them from fluent or typically disfluent speech.

The current gold-standard of articulatory research uses electromagnetic articulography (van Lieshout & Moussa, 2000; Yunusova, Green, & Mefferd, 2009), where the data informs about individual points on the tongue. Coils are placed on the tongue surface and their movement through space and time is measured. The placement of the coils should be tailored to the individual's articulator size and shape, which is difficult given that the coils are usually placed without full articulatory information. Ultrasound tongue imaging has the advantage that it captures almost the entire tongue surface, which can be observed moving in space and time (Stone, 1997, 2004). Much of previous ultrasound research has employed a static approach, where the tongue surface is investigated at time stamps relative to the acoustic signal. A transition deficit as proposed by Wingate (Wingate, 1969b, 1988), however, would not be accessible when only considering the tongue configuration at certain points in time.

Instead, we employ a novel approach where we bring together advantages of techniques used in electromagnetic articulography and ultrasound tongue imaging. We use the movement of almost the entire tongue surface to localise the region of interest on the tongue surface. The temporal and spatial movement of that region is then extracted for the CV transition.

This thesis seeks to explore the question of whether it might be necessary to extend the study of stammering beyond perceptually identifiable instances of disfluency to explore the perceptually fluent production of people who stammer. To compare information from the surface level as well as the underlying source level, we conducted this study using articulatory measures in addition to traditional acoustic measures. Building on Wingate's Fault-Line hypothesis, we employ a novel approach to obtain kinematic information using articulatory data to investigate consonant-vowel transitions. We employ ultrasound tongue imaging data to explore the lingual kinematics in the transition between consonant and the subsequent vowel as described by Heyde and colleagues (Heyde, Scobbie, Lickley, & Drake, 2016). High resolution ultrasound tongue imaging data was obtained to inform about the midsagittal tongue contour over time. Additionally, established acoustic measures were applied to investigate motor timing and control strategies in PWS and PNS.

1.3 Models of Typical Speech Production

To consider further the theoretical framework, which motivated the decision to study consonant-vowel transitions in the fluent speech of people who stammer, it is important first to understand how typical speech is produced: The realisation of speech is a highly complex process. Since the 1960s, researchers have proposed multiple models trying to explain the mapping between the cognitive representation of phonemes and the respective articulatory gesture. The problem is that a cognitive representation does not map directly onto a linguistic representation and the same is true for the mapping between articulatory

realisation and the acoustic output. As mentioned previously, there is no direct one-to-one translation between these.

Languages and dialects differ in their representations and their realisations. Linguistic targets are arbitrary and as such speakers of different languages have different words for the same entity. What is referred to as a 'tree' by an English speaker, is called 'Baum' by a German speaker – both referring to the same entity, a plant with a stem, branches and typically with leaves. While linguistic targets differ between languages, they all draw on common building blocks, namely CV syllables (Krakow, 1999). In MacNeilage and Davis's Frame and Content model (B. L. Davis & MacNeilage, 1995), these are the starting points for children as they tend to produce CV syllables in canonical babbling. The way the acoustic target maps onto the articulatory representation, however, is highly complex.

More importantly, a single acoustic target does not map onto just one articulatory representation. Instead, various articulatory realisations are possible. Speech is adjusted to the situation it takes place in but also to the speaker it is produced by. If, for example, we find ourselves in a loud environment we may need to speak up ('The Lombard effect': Šimko & Beňuš, 2016). Or in situations where messages need to be delivered quickly, the speech rate needs to be increased, which might not leave enough time to fully reach each articulatory target (Loucks & De Nil, 2012). The physiognomy of every speaker differs and so does the execution of his or her speech. Vocal tracts are individual to every speaker (Fuchs, Perrier, Geng, & Mooshammer, 2006; Rudy & Yunusova, 2013). Palates differ in height and shape, tongues differ in size. Even the speaker's command of his tongue differs. A certain degree of variation in the execution of an articulatory configuration is therefore inevitable. Speakers need to adapt their articulatory target to their oral cavity as well as to their articulatory control. This falls under the term 'motor equivalence' (Tasko & McClean, 2004) which expresses that articulatory configurations may differ while still yielding the same acoustic targets.

But articulatory targets are not only produced differently because of extra-linguistic reasons such as physiognomy or control. Intra-linguistic dependencies also affect the way speakers articulate. In the production of strings of speech, neighbouring segments of speech affect each other. The execution of one segment of speech varies as a function of its neighbouring sounds, owing to carry-over and anticipatory coarticulatory effects of those neighbouring sounds (Barbier, Perrier, Ménard, Tiede, & Perkell, 2013; Recasens, 2002; Zharkova, Gibbon, & Hardcastle, 2015). Hence, articulatory targets differ depending on the individual and the context in which these targets are produced. Anticipatory coarticulation is typically found greater when compared with perseverative coarticulation with degrees of coarticulation differing as a function of place of articulation (Krull, 1989a).

While all these extra-linguistic as well as intra-linguistic factors influence the way we produce speech, the variation with which we produce speech is structured and not without limits. The variability with which articulatory movements can be executed is limited by two main factors: for one, there is the anatomic limitation allowing for / favouring certain articulatory configurations and then there is the ability of the listener. Because the function of speech is to convey a message, it needs to be produced in a way that the acoustic output can be related to the intended message by the interlocutor. Both the physical and the linguistic factors underlie the concept of 'degrees of freedom' (Perrier et al., 2007).

In the following section, we lay out theories / models on speech planning (see 1.3.1) and speech production (see 1.3.2) in typical speech, which form the basis for an understanding of the fluent speech of people who stammer.

1.3.1 Speech Planning: From Conceptualisation to Articulation

The term 'speech production' comprises different stages starting from the conceptualisation of a message to the execution stage where words and full sentences are articulated. Each stage is highly complex and to produce speech multiple complex processes are tightly coupled to achieve smooth and efficient

movements (Šimko & Cummins, 2011). Speech production is usually divided into three stages, starting with a conceptualisation stage where the intended non-verbal message relating to a concept is formulated. In a second stage, the non-verbal conceptual message is linked to the different levels of linguistic means (i.e., syntactic, morphologic, phonologic, and phonetic encoding) that are required to express the intended message. The sentence structure consists of slots that are filled with words that are made up of lemmas that are linked to phonetic units. Phonetic encoding involves the activation of articulation. At this point articulatory gestures are activated relative to a speech sound, which is then used as input to the third stage, the execution of the articulatory score where the actual speech sound is produced.

1.3.1.1 Serial Models of Speech Production

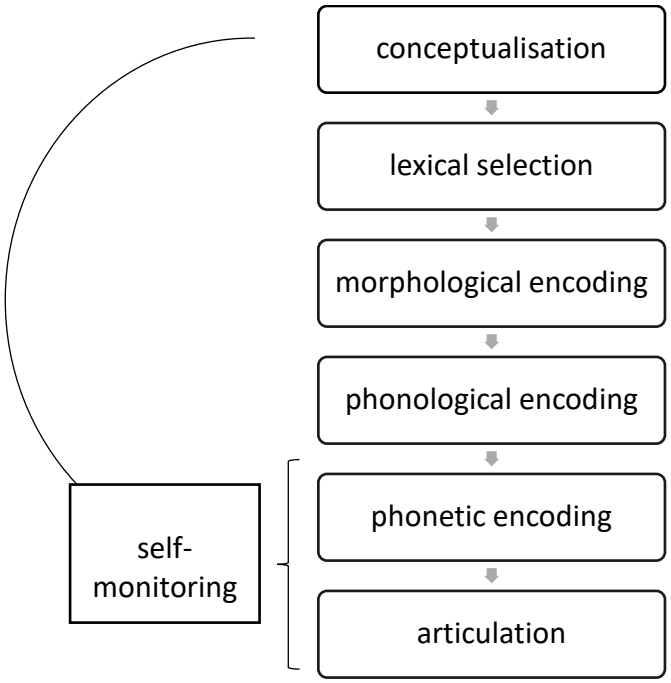


Figure 3 *Serial model of speech processing based on Levelt et al.* (Levelt, Roelofs, & Meyer, 1999)

Initially researchers proposed serial processing models (see Figure 3 *Serial model of speech processing based on Levelt et al.* (Levelt, Roelofs, & Meyer, 1999)Figure

3)where speech was produced as a series of sequential stages (Fromkin, 1971; Garrett, 1975; Levelt, 1992; Levelt et al., 1999), with Levelt's model of speech production being amongst the most influential. Serial models propose several processing stages that are executed hierarchically starting with the larger semantic units and then moving towards the smaller phonological units – each having a characteristic output representation.

The unidirectional serial nature of speech processing was soon questioned. Regarding the unidirectional nature of speech processing, Levelt (Levelt, 1983, 1984) was one of the first to introduce a self-monitoring process where speakers would monitor their speech and repair errors overtly (at the phonetic / articulatory level) as well as covertly (at an earlier phonological level). This way he accounts for overt and covert speech errors as well as for hesitations and pauses. Levelt later also questioned the serial nature of speech production by posing the question whether these stages are indeed purely sequential or whether they may be overlapping. In doing so, he opened the door for the next generation of parallel speech processing models.

1.3.1.2 Parallel Models of Speech Production

Dell et al. (Dell, Chang, & Griffin, 1999) formulated a model in which different levels function like an interconnected network where the sentence and articulatory levels can feed back into the conceptualisation level. For their model Dell et al. adopted the 'Connectionist Model' (McClelland & Rumelhart, 1981; Munakata & McClelland, 2003; Rumelhart, 1998), which explains lexical access through spreading activation of shared semantic meaning as well as the shared phonological segments of a word. Units of speech are represented as connected nodes that can interact in any direction. Nodes on the different levels of representation are activated simultaneously and compete for the highest activation. The node with the highest activation is then selected.

According to parallel models, speech errors occur respective to the level of representation of the node that has erroneously received the most activation.

Speech errors may, however, even occur when the correct node receives highest activation and is selected. Competing nodes receive partial activation which influences the word that will be produced. The co-activation of nodes for the words ‘strong’, ‘tough’ and ‘rough’ may lead the speaker to produce ‘strougher’ when he actually intends to produce ‘tougher’. This way spreading activation models (as opposed to serial models) account for substitutions and blends of phonemes (word blend) or entire words (phrase blend; showing that the scope of speech production planning is bigger than the word). Note that there is a crucial difference in perception and production. The latter operates in units of speech imposed on a known word. Perception, in contrast, must slice continuous speech into competing lexicalisations.

Parallel models of speech processing also account for intrusion errors on the articulatory level as described by Pouplier (Pouplier, 2007; Pouplier & Goldstein, 2010) which we will discuss briefly in section 1.3.3. Considering CV transitions in contrast, the question arises whether breakdowns in fluency occur as an issue at the level of phonetic encoding (as proposed by Wingate) or between phonological and phonetic encoding.

1.3.2 Speech Production: From Motor Planning to Execution

In acoustic phonetics the model of the speech execution stage consists of two components with (a) sound coming from the noise source at the larynx and entering the vocal tract and (b) the filtering and modifying of that airstream in the vocal tract (Bouchard, Mesgarani, Johnson, & Chang, 2013; Fant, 1970). In articulatory phonetics, this execution stage is broken down into smaller sub-stages where the lungs, glottis and larynx in concert with a combination of configurations of tongue, lips and jaw translate these abstract gestures into speech.

As part of the speech execution stage, air is pushed up the vocal tract starting from the lungs through the glottis where the opening / closing of the glottis differentiates between voiced (i.e., vibrating vocal folds) and voiceless sounds. With the glottis

open and the vocal folds drawn apart, the air can travel through without causing the vocal folds to vibrate, resulting in voiceless phone. If, however, the glottis is closed, the airstream coming from the lungs presses against and excites the vocal folds, which initiates a recurring cycle of vibrating vocal folds. Starting from the closed glottis the airstream presses against the glottis and the increasing pressure causes the glottis to open. This is when the air can escape and the vocal folds return to their closed starting position. Cycling through the opening and closing of the vocal folds, we perceive vocal fold vibration required for both the production of vowels and voiced consonants.

The timing of turning voicing on and off can be important linguistically. If we think of, for example, the onset of voicing in a vowel after an initial voiceless consonant, the turning on of voicing needs to be coordinated with other active articulators, such as the tongue and lower lip. Independent of whether the sound produced is voiced or voiceless, once the air stream has passed through the larynx the sound is filtered / modified in the vocal tract (Fant, 1970). For consonants, traditionally, two main components of vocal tract modification are distinguished, i.e., manner of articulation and place of articulation. The expression 'manner of articulation' refers primarily to consonants where the airstream is either fully or in parts blocked by the tongue or lips forming, for example, fricative or stop consonants, which we will now turn to. The latter, 'place of articulation' relates to where in the vocal tract the full or partial closure occurs. 'Place of articulation' is described as a function of the primary articulator involved in the perturbation of the airstream, i.e., labial, dental, or lingual. For lingual consonants, the closure location is typically described in relation to the passive articulator resulting in consonants that are defined as, for example, dental, alveolar, or velar. For the modification of vowels, tongue body height is changed, and the vocal tract size adjusted by spreading or rounding the lips. Following this scheme, a large repertoire of separate acoustic targets can be realised. To produce strings of speech more factors play a role in the translation of intended acoustic strings of targets into an appropriate articulatory realisation, which we will not be able to cover in this thesis due to limitations in scope.

In order to understand the complexity of transitioning between consonant and vowel, it is necessary to look at models of speech production and at how they account for the most basic form of syllables, i.e., CV syllables, in typical speech.

1.3.3 Syllable Structure

Traditionally, syllables are considered the basic unit of speech production. Pike and Pike (1947) introduced the internal constituent structure suggesting that three main constituents underlie the basic syllable structure (see Table 1). These syllable constituents include onset, peak / nucleus, and coda. In 1972 Newman (Newman, 1972) introduced rhyme as the head constituent sharing the same syllable node with the onset and branching into nucleus and coda (see Figure 4).

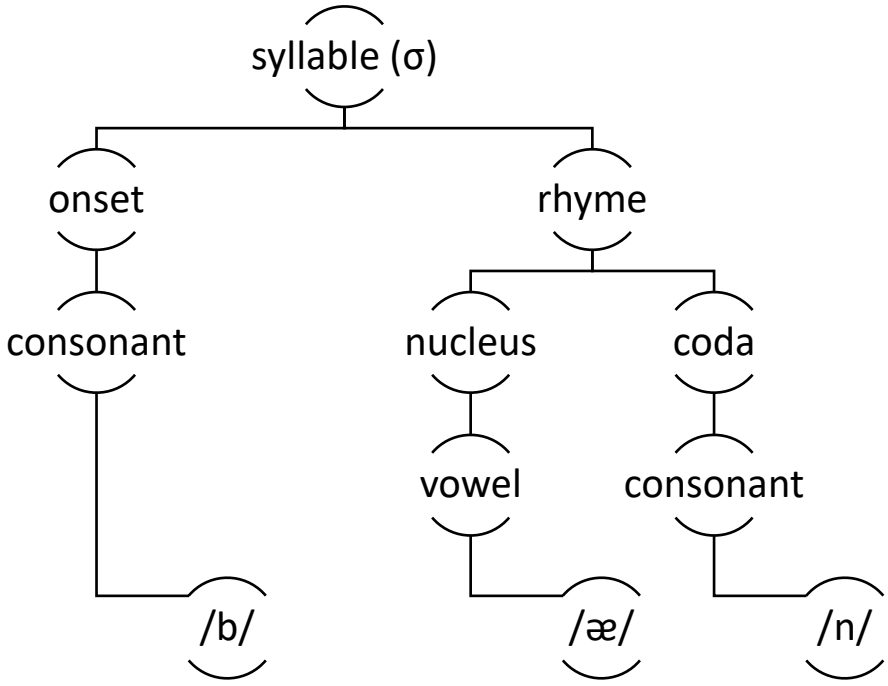


Figure 4 *Graphic representation of a CVC syllable including onset and rhyme (consisting of nucleus and coda)*

In the given example for the word ‘ban’ (Figure 4), the three syllable constituents onset, nucleus and coda are each occupied by one segment. When we look at other words, however, we see that syllable constituents can take more than one segment

and that onset and coda are not obligatory to the syllable while nucleus is (see Table 1).

Table 1 *Syllable Structure according to Pike and Pike (1947)*

	onset	nucleus	coda
ban	/b/	/æ/	/n/
glue	/gl/	/u:/	Ø
end	Ø	/ɛ/	/nd/
eye	Ø	/aɪ/	Ø

1.3.3.1 Articulatory Models

Articulatory mechanisms have been employed to account for the temporal coupling of these syllable constituents that would hold despite changes in speaking rate or prosody. The first dynamic models for articulation were presented as early as the 1960s. Henke (Henke, 1966), for example, proposed a dynamic articulatory model to produce speech using computer simulation. In his model, Henke used segmental input where each segment consisted of a set of phonetic features defining the state of the vocal tract. Via operators, the state of the model was then modified as to achieve articulatory targets for vowels and stop consonants. This dynamic articulatory model already incorporated coarticulatory effects resulting from anticipatory processes (Barbier, Perrier, Ménard, Tiede, et al., 2013). Among the most influential models were the Task Dynamic Model and Articulatory Phonology, which we will briefly describe in the following sections.

1.3.3.1.1 Task Dynamic Model

The Task Dynamic model (Kelso, Saltzman, & Tuller, 1986; Saltzman & Kelso, 1987; Saltzman & Munhall, 1989) was originally embedded in the field of movement sciences. It changes the focus from the target (i.e., intended speech sound) away to the movements required to achieve the target. ‘Gestures’ capture the idea of a

combination of movement patterns that within a task space is required to achieve a specified target. When applying the Task Dynamic Model to speech, the basic assumption is that the same principles that underlie any skilled motor action also apply to the control and coordination of speech movements.

Discrete movements are broken down into several independent tasks. For the case of articulatory movements, typically, the tasks are derived from vocal tract constrictions as they define the parameters of aperture of velopharynx and glottis, degree of lip protrusion and lip aperture as well as tongue constriction and location where each articulator can contribute to multiple parameters simultaneously. If different trajectories of articulators can achieve a single target this is referred to as 'motor equivalence' (Guenther, 1994). For each trajectory, the parameters of invariance and variability are adjusted. Articulatory movements are modelled as a mass attached to a spring (representing the changes in the tract variable) and a damper to avoid oscillatory movement. To generate multiple gestures, simultaneous commands are sent to the articulators, which accounts for coarticulation where adjacent phonemes are blended (Kühnert & Nolan, 1999). Following Wingate's argumentation that stammering is caused by a transition deficit, we would expect an imbalance in the parameters of invariance and variability.

1.3.3.1.2 Articulatory Phonology

As can be seen from the above model descriptions, traditional approaches to modelling speech production have considered two independent structures: the physical articulatory and the cognitive linguistic structure. While the former is constantly moving in time and space, the latter appears to be more rigid with a defined and therefore limited inventory, "where the relation between them was generally not an intrinsic part of either description. From this perspective, a complete picture requires 'translating' between the intrinsically incommensurate domains (as argued by Fowler, Rubin, Remez, & Turvey, 1980)" (Browman & Goldstein, 1995, p. 176). Articulatory Phonology (Browman & Goldstein, 1989,

1990, 1992a, 1995; Browman, Goldstein, & Ohala, 1986) begins with the very different assumption that these apparently different domains are, in fact, the low and high dimensional descriptions of a single (complex) system. Crucial to this approach is identification of phonological units with dynamically specified units of articulatory action, called gestures. Browman and Goldstein follow Fowler in breaking the boundary between phonology and phonetics. The idea of gestures is adopted in that they are seen to function on both the phonological and the phonetic level. Articulatory Phonology further includes the task dynamic principles introduced by Saltzman and Kelso (Kelso et al., 1986; Saltzman & Kelso, 1987) to speech production and the possibilities of simultaneously produced gestures.

The Articulatory Phonology model proposes that phonetics and phonology exist within the same system where "phonology is a set of relations among physically real events, a characterization of the systems and patterns that these events, the gestures, enter into" (Browman & Goldstein, 1992a). Gestures are defined as "characterizations of discrete, physically real events". They are "basic [phonological] units of contrast among lexical items as well as units of [phonetic] articulatory action [that] can be used to capture both categorical and gradient information" (Browman & Goldstein, 1992a).

The abstract phonological system is constrained by the articulatory system of phonetics, thereby aligning the phonological and the phonetic units. Five gestural families (velopharyngeal, glottal and labial aperture, tongue tip (TT) and tongue body (TB)) are defined. Browman and Goldstein diffuse the boundaries between what is traditionally seen as two levels with their idea of articulatory gestures representing 'units of combination and contrast' (phonological) as well as 'units of action' (phonetic).

Gestures are defined in terms of the degree and location of the constriction and the stiffness required to achieve the constriction. Gestural scores capture the temporal (sequencing) and dynamic (stiffness) parameters for the sequence of gestures. These scores function as abstract representation of articulatory scores (e.g., lip

closure for /p/) that are defined independently from surrounding context so that the gestural score for /p/ would be the same in, for example, sequences like /ipi/ or /apa/ (Recasens & Espinosa, 2009).

Using the principles of Task Dynamics (Saltzman, 1986), articulatory movements are described as dynamic tasks (instead of static targets) that need to be performed. They carry information about the target, but also information on how the target is achieved. Gestures are dynamic and articulatorily invariant that carry temporal specifications that are not altered in context. Tract variables (e.g., lip aperture, tongue tip constriction location and degree) are used to capture the constriction location and dimension.

Table 2 *Gestural scores (adopted from Browman & Goldstein, 1990)*

	Tract variable	Articulators involved
LP	lip protrusion	upper and lower lips, jaw
LA	lip aperture	upper and lower lips, jaw
TTCL	tongue-tip constriction location	tongue-tip, tongue-body, jaw
TTCD	tongue-tip constriction degree	tongue-tip, tongue-body, jaw
TBCL	tongue-body constriction location	tongue-body, jaw
TBCD	tongue-body constriction degree	tongue-body, jaw
VEL	velic aperture	velum
GLO	glottal aperture	glottis

The gestural score (see Table 2) captures temporal intervals for the distinct vocal tract actions. The tract variables are temporally arranged indicating the duration for each gesture as well as the temporal overlap between gestures.

In the example in Figure 5 we see the gestural scores for the two words ‘bad’ (in the upper panel) and ‘ban’ (in the lower panel). The gestural scores are very similar with the only difference being the presence (in ‘ban’) or absence (in ‘bad’) of velic lowering. This example shows that the presence or absence of gestures carries important information.

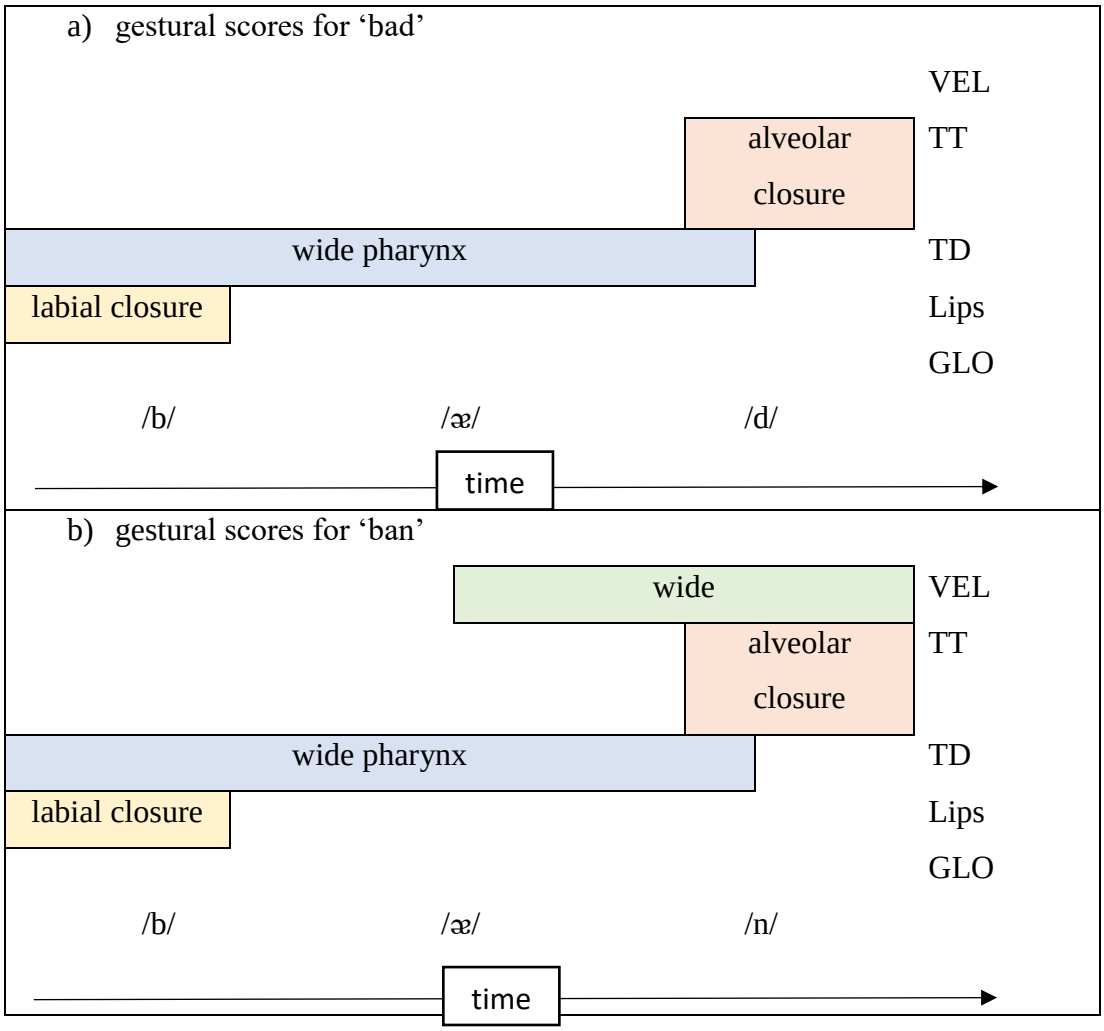


Figure 5 *Gestural scores for the words ‘bad’ (upper panel a) and ‘ban’ (lower panel b) indicating vocal tract constrictions for the five gestural families (y-axis) over time (x-axis) (adopted from Goldstein, Nam, Saltzman, & Chitoran, 2009)*.

But it is not just the presence or absence of gestures that is important. Equally important is the phasing of these gestures. In the example in Figure 6 the same gestures are active for the words ‘mad’ and ‘ban’ with the only difference being the

temporal relationship between the gesture of velic widening. To account for the intrinsic temporal relationships of these speech units, Goldstein et al. (Goldstein, Chitoran, & Selkirk, 2007) have proposed the Coupled Oscillator Model which we will discuss in the following section.

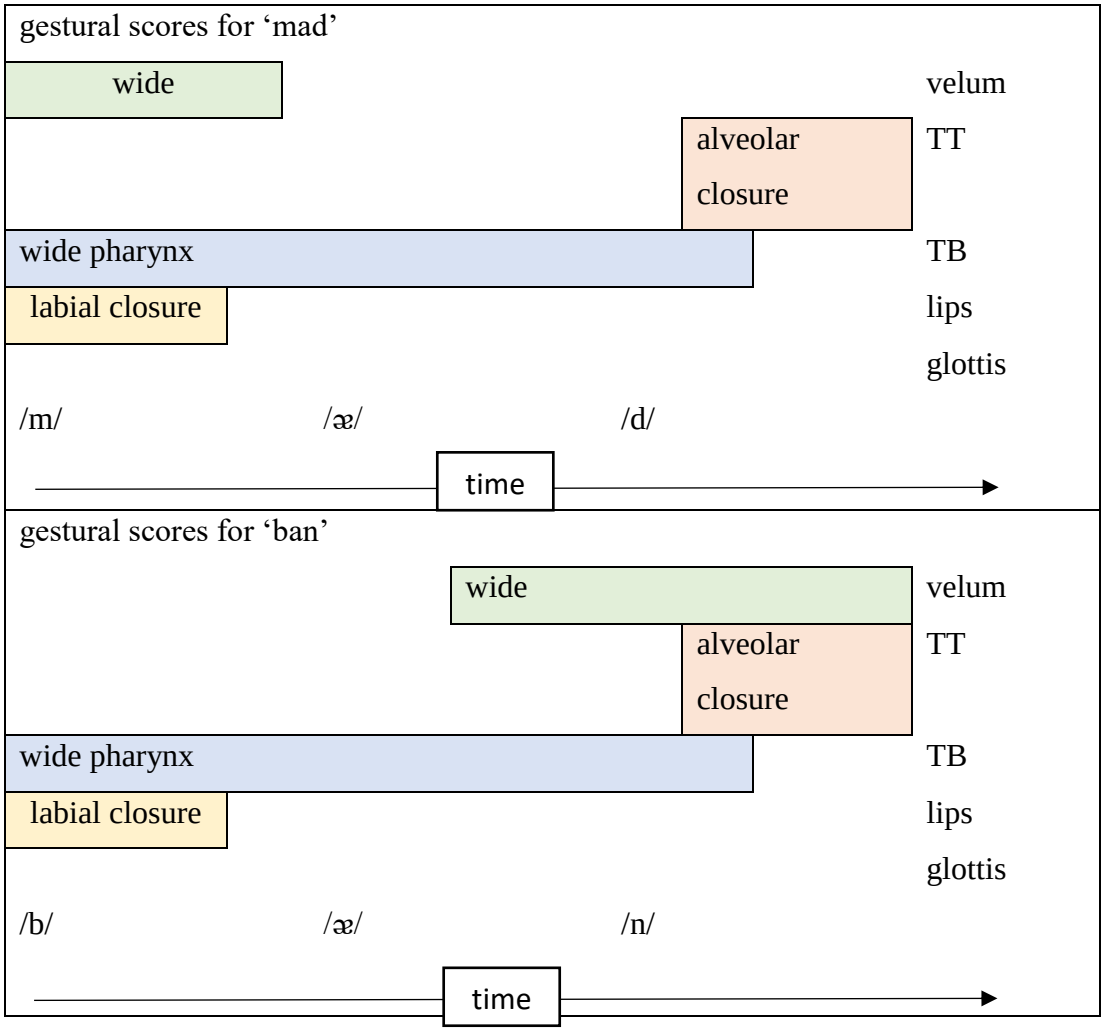


Figure 6 *Gestural scores for the words 'mad' (upper panel a) and 'ban' (lower panel b) indicating vocal tract constrictions for the five gestural families (y-axis) over time (x-axis) (adopted from Goldstein et al., 2009).*

1.3.3.1.3 The Coupled Oscillator Model

The Coupled Oscillator Model (Goldstein, Chitoran, et al., 2007; Goldstein et al., 2009) provides a framework for the intrinsic timing proposed by both Task Dynamics (1.3.3.1.1) and Articulatory Phonology (1.3.3.1.2). The model assumes

that timing is oscillator based. Gestures are associated with oscillators to account for the relative temporal relationship of these gestures. Constriction gestures are associated with oscillators which are coupled in pairs. The intrinsic coupling of gestures makes their relative timing independent from extrinsic factors such as, for example, the overall speaking rate or prosody.

The Coupled Oscillator Model builds on the framework of Articulatory Phonology, also using gestures as the basic units of speech production. In this model, the timing between the individual gestures is related to mechanisms of gestural phasing.

The model defines two types of phasing:

- First, there is in-phase coupling. Onset (consonant) gestures and the gesture for the following vowel in syllable nucleus position are hypothesised to be coupled in-phase as both gestures begin synchronously.

- Second, there is anti-phase coupling, which is claimed to apply to nucleus (vowel) and coda (consonant) gestures because their gestures begin with a time lag. The consonant gesture begins roughly when the gesture for the vowel ends.

Gestures that are coupled in-phase are said to be most stable as they are acquired earlier and planned faster (Saltzman, Tyrone, & Goldstein, 2000). Relating this oscillator model back to the syllable structure, CV syllables are said to be in-phase, whereas VC syllables are produced anti-phase. Goldstein et al. (Goldstein et al., 2009) claim that the production of in-phase CV syllables is less problematic and more stable when compared to that of VC syllables.

Evidence for gestures as the basic units of speech production comes from articulatory studies on speech errors. Goldstein et al. (Goldstein, Pouplier, Chen, Saltzman, & Byrd, 2007) showed that speech errors obey phonological rules, which makes them an important source for studying the nature of speech units. Errors showed a general preference of speakers for in-phase productions over anti-phase

productions confirming the coupling of gestures as proposed in the Coupled Oscillator Model. Participants were asked to produce CVC syllables with alternating onset consonants. Different speech rates were elicited using a metronome. Materials included the sequences 'top cop' and 'cop top', while control materials did not alter in onset consonants (i.e., 'top top' or 'cop cop'). In the control material for 'top top', each cycle of the tongue tip (for /t/) is associated with a cycle of the lip (for /p/). The same is true for the tongue body (for /k/) and the lip (for /p/) in 'cop cop'. In both cases, the cycles for tongue and lip constrictors are locked at 1:1 frequency (in-phase). Haken et al. (Haken, Peper, Beek, & Daffertshofer, 1996) have shown that 1:1 frequency locking is the most stable of frequency lockings. This is supported when looking at the control material where participants did not produce errors. The alternating syllable onset material 'top cop' or 'cop top' results in 1:2 (anti-phase) frequency locking for /t/ and /p/ cycles as well as for /k/ and /p/ cycles, which elicited more errors when compared to the control material with 1:1 frequency locking.

The error patterns further revealed a preference for intrusion rather than substitution. With increasing speed, speakers show a tendency to transition from the more complex 1:2 frequency locking to the simpler 1:1 frequency locking, whereby gestures for both /t/ (tongue tip) and /k/ (tongue dorsum) syllable onsets are produced simultaneously before /p/ coda is reached. Using acoustic information, these errors were often either acoustically misinterpreted as reductions and substitutions of the original segment or not perceived due to their low magnitude. Using articulatory imaging, Pouplier and Goldstein (Goldstein, Pouplier, et al., 2007; Pouplier, 2007) have demonstrated a discrepancy between information available from acoustic and articulatory data. Articulatory data revealed a bias towards gestural intrusion errors indicating coactivation of gestures as predicted by parallel models of speech production (see 1.3.1.2).

The intrinsic approach to timing is promising as it uses the overlap of gestures also explaining the vocal tract changes over time. Further, timing specifications are

derived from the phonological plan, which makes a phonetic plan for timing dispensable.

For CV transitions, the coupled oscillator model predicts tight coupling of the gestures for the consonant and for the vowel. For speakers with intact motor control, one would therefore expect consistent coupling over multiple repetitions of the same token. In fluent as well as disfluent productions, PWS have the same gestural score as typical speakers. Regarding the consistency of their gestural coupling, however, we might expect to find larger variation for PWS.

1.3.3.2 Importance of Coarticulation

In moving away from the target towards the movement that is required to achieve the target, coarticulation has become an increasingly central question to speech research (Volenec, 2015). Because transitions between targets are central to the current study, we will briefly explore how coarticulation affects these consonant-vowel transitions in typical speech.

Coarticulation affects both the target locations of discrete gestures and the transition between them as a function of neighbouring segments. The target location of one gesture shifts towards that of the neighbouring gesture (Öhman, 1966). For example, the velar closure location differs with respect to different vowel environments. As Frisch and Wodzinski (Frisch & Wodzinski, 2016) have demonstrated, the target location of the velar constriction is fronted in the context of front vowel /i/ when compared to the context of the open vowel /ɔ/. Another example of coarticulation is that of lip rounding. In the sound sequence /b + u/, for example, the lip rounding required for the round vowel /u/ starts already during the preceding consonant.

Even prior to the outlining of Articulatory Phonology, researchers had suggested that coarticulation might reflect the structure of programming units in the production of speech. Kozhenikov and Christovich (Kozhevnikov & Christovich,

1965) have stated that CV is the basic unit of speech, thereby supporting the idea of speech being organised in syllabic units.

Kent and Minifie provide a general definition of coarticulation where the concept of coarticulation is captured in two arguments "(1) discrete and invariant units serving as input to the system of motor control, and (2) an eventual obscuration of the boundaries between units at the articulatory and acoustic levels" (Kent & Minifie, 1977). Kent and Minifie add that the obscuration of boundaries happens two ways a) via anticipatory coarticulation, also referred to as right-to-left, regressive or forward coarticulation and b) via backward / perseverative coarticulatory effects. Over time a more distinct map of coarticulation theories has formed (Farnetani & Recasens, 1999; Kent & Minifie, 1977; Volenec, 2015; Zharkova, Hewlett, Hardcastle, & Lickley, 2014).

One approach to explain coarticulation has been the target-oriented feature based approach (Henke, 1966; Moll & Daniloff, 1971) where phonetic units are ascribed a valence (+/-) for the different features (e.g., voiced, lip protrusion, velum raising, etc.). The feature-based models claim that phonetic representations are specified regarding their articulation, but also coarticulation. To account for coarticulation, researchers supporting the feature-based approach have applied the principle of feature value compatibility. This approach suggests that adjacent feature values cannot be contradictory. In cases where features are contradictory, coarticulatory effects of adjacent segments will be essential to manage the transition between them. While feature-based models may explain anticipatory as well as perseverative coarticulatory effects, their limitation is that these feature-based models can only account for directly adjacent segments. This theory, however, fails to account for findings of coarticulatory effects from / to non-adjacent segments (Öhman, 1966; Stetson, 1928).

Hierarchical models, in contrast, account for coarticulatory effects of adjacent as well as non-adjacent segments. In these models the different levels of speech organisation are emphasised: "the features that constitute the segments are

organized in production into syllabic bundles, each consisting of overlapped and largely independent articulatory components" (Liberman, 1970, p. 313). The models therefore allow for interaction between elements on the same level (phones, features and muscle gestures, but also on the syllabic level) as well as between levels.

1.3.3.2.1 Coarticulation Affecting Articulation

Different explanations have been proposed for the occurrence of coarticulation. One of them suggested that inertia would partially account for these effects of coarticulation. Daniloff and Hammarberg, for example, put forward a purely physiological account where the intended form does not match the capacities available for execution (Daniloff & Hammarberg, 1973; Björn Lindblom, 1963). In their theory they see coarticulation as resulting from the limitation of the level of acceleration that can be achieved by articulators. Articulators need to transition between the discrete units of speech, i.e., phonemes and in fact, the transition itself often may take up more time than the time the articulator spends at the actual target. In some cases, the transition from one articulatory target position to the subsequent might even take too long to fully execute the target before moving on to the next.

Using a similar line of argumentation, Lindblom formulates the 'target undershoot model' claiming that the articulators' biomechanical properties inhibit the full execution of an articulatory command before moving on to the next. The limited speed with which articulatory gestures can be executed leads to a carry-over effect whereby commands are extended into the next gesture.

Another model by Lindblom is referred to as the Speech Economy Model (Lindblom, 1983, 1990). In contrast to the Target Undershoot Model it targets a higher communicative level. This model integrates two communication-based principles that the speaker must adhere to. First, the listener-oriented principle that is required for the achievement of successful communication. The contrast between

segments must be maintained to enable the interlocutor to follow and segment the speech signal. In contrast stands the speaker-oriented principle where the speaker attempts to produce speech with as little effort, while producing the required perceptual contrast to convey his message and keep the conversation going. These two principles demonstrate that both social factors and lexical factors matter. Though the phonetic balance between features may differ, they may yield the same articulatory-acoustic results (Lawson, Scobbie, & Stuart-Smith, 2014 on socio-phonetic adaptation; Mayo & Turk, 2004; Nittrouer, 2006 on developmental change).

Inertia and speech economy more generally may account for coarticulation in parts. The carryover effect in the target undershoot model accounts for coarticulatory patterns of neighbouring phonemes. Neither, however, explain the systematicity underlying, for example, the coarticulatory resistance of certain phonemes when compared to others (Recasens, 1985).

1.3.3.2.2 Coarticulation and Gestures

A different approach to coarticulation was introduced by Fowler (Fowler, 1980) who formulated the Coproduction Theory where both segments maintain their inherent qualities and coarticulation is the result of the two segments being produced in overlap and blend as can be seen in Figure 7.

For Fowler, phonological and phonetic units are both specifying an articulatory gesture that is context independent. While the phonological representations have a specified temporal dimension, these are not specified for gestures. Gestures usually take longer than the associated segments. Segments can be conceived as non-overlapping entities, which can be easily defined in the acoustic signal. The associated longer gestures, in contrast, are by necessity overlapping, which results in coarticulation, which can also be observed in the acoustic signal. While the fundamental ways of looking at speech using the acoustic signal are valid, the

investigation of the articulatory signal does provide the direct access to articulation that necessarily underlies speech production.

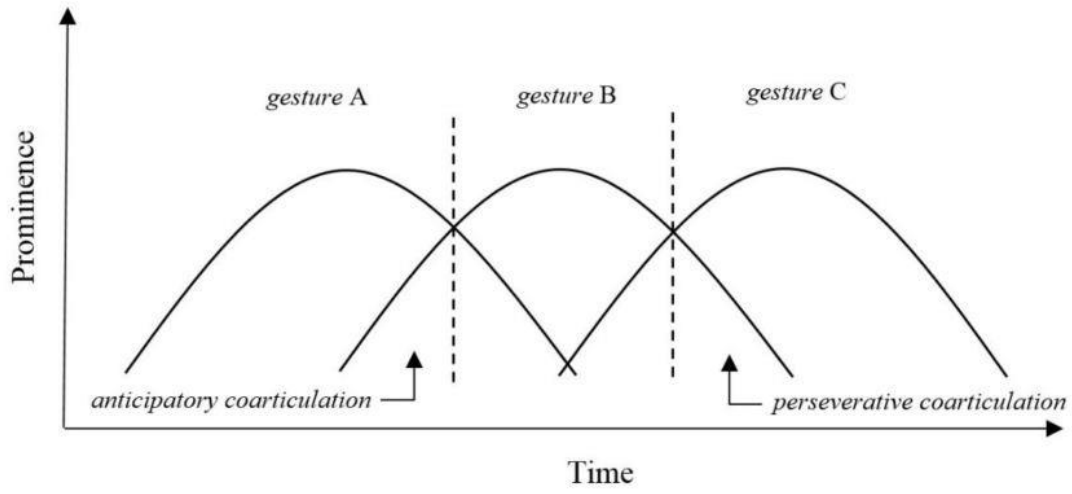


Figure 7 *Schematic representation of three overlapping gestures and the anticipatory as well as perseverative / carryover coarticulation (Fowler & Saltzman, 1993).*

Combining the gestural scores proposed in Articulatory Phonology (Browman & Goldstein, 1992a) and the Coupled Oscillator Model (see 1.3.3.1.3), a different approach to coarticulation becomes available where coarticulation is part of speech planning. Speech sounds are not just stringed together with coarticulation falling out of them. Instead, gestural scores demonstrate that gestures typically occur in overlap. Browman and Goldstein explain coarticulation as resulting from the coproduction and blending of gestures whenever articulators are shared (Fowler & Saltzman, 1993; Meyer & Gordon, 1985). The relative timing of these gestural scores (overlap of two or more gestures) accounts for coarticulatory patterns. The Oscillator Model claims that gestures are coupled in pairs, which suggests that coarticulation does not happen by chance but underlies rules that specify how gestures are coupled.

Assuming the coupling of gestures, the simplest forms of coarticulation can already be observed at the level of basic consonant-to-vowel transitions. It is often claimed that syllables are the basic physiologic unit in speech production (Krakow, 1999), accounting for the coarticulatory effects described above. Coarticulation, therefore, is evidence for speech planning happening at a higher level than the phoneme level. This is supported by coarticulatory patterns where coarticulation is greater within syllables than across syllables (Chang, Ohde, & Conture, 2002; Daniloff & Moll, 1968; Xu, 2017), emphasising the importance for articulatory patterning in speech.

Being a motor skill, linguistic coarticulation in speech is a learned behaviour where coupling increases with maturity. A pattern of decreasing gestural variability with increasing age and maturity is often found in research more generally (Hourcade, Bederson, Druin, & Guimbretière, 2004; Jansen-Osmann, Richter, Konczak, & Kalveram, 2002; Lambert & Bard, 2005) but also in research on speech motor skills (A. Smith & Goffman, 1998) and research on coarticulatory patterns more particularly (Zharkova, Hewlett, & Hardcastle, 2011). For stammering, what this might imply it that poorer control of intergestural timing could result in increased variability, which we will explore throughout this thesis.

1.3.4 Summary

Based on Articulatory Phonology we understand typical speech to consist of sounds which are acoustic targets executed via motor commands. These motor commands specify articulatory target locations in the vocal tract. For each sound the articulators are positioned to achieve the required degrees of constriction in the vocal tract. Different movement patterns correspond to articulatory gestures.

In running speech, discrete sounds of speech are stringed together. To execute these strings of sound, the speech organs must be well orchestrated to coordinate the different adjacent gestures in a timely manner. The timing of adjacent gestures is controlled via intrinsic coupling mechanisms that are expressed on the syllable level through coarticulation.

1.4 Stammered Speech

Speech errors are amongst the phenomena that motivate the model of Articulatory Phonology. Their properties reveal properties of speech production, monitoring and control. Disfluencies in stammered speech may be considered as a type of speech error. In the following section, we will briefly review stammering theories that construct stammering as a language or speech impairment.

Stammering is typically characterised by a relapsing-remitting, often situation-specific pattern of symptoms – primarily involuntary disruptions in the smooth flow of speech. The most common symptoms are described in terms of their acoustic consequences, labelled as blocks, prolongations and repetitions (Guitar & Belin-Frost, 1998, p. 213). The three symptoms have in common the successful achievement of the initial consonant. For the purpose of this thesis, achieving the point of maximal closure is treated as achievement of a consonant posture. All three symptoms further share the difficulty when moving into the subsequent vowel. The difficulty may be two-fold where PWS struggle to either move away from the consonant or to move into the vowel which we will address in the discussion section. Wingate represents the latter view in his Fault-Line Hypothesis (Wingate, 1969b), which we will discuss in section 1.4.2.1.

It should be noted that disfluencies are not exclusive to consonant-initial syllables. Stammering is also frequently observed on vowel-initial syllables, which are in fact often glottal-stop initial syllables. While it will not be covered in this study, it is worth investigating whether glottal stops may have a similar status as consonants in the context of stammering.

The successful achievement of the initial consonant and the difficulty transitioning into the subsequent vowel appears to be a common feature. This raises the question whether disfluencies should be treated in these three separate categories of blocks, prolongations and repetitions or whether they are more alike in that they all constitute manifestations of problematic transitions. Their nature could depend

entirely on the manner of articulation and the severity of the breakdown in fluency. Some researchers claim blocks to be the characteristic type of disfluency in the speech of PWS where repetitions and prolongations constitute coping mechanisms. Other researchers see a relation between repetitions and blocks where the latter represents a more severe breakdown in fluency. The burst in a plosive consonant is rapid in nature and is therefore unlikely to be extended, which renders prolongations unsuitable for plosive consonants. For the same reasons, repetitions are more likely to occur on plosive consonants.

When we think of stammering, most people think of persistent developmental stammering (Prasse & Kikano, 2008). Developmental stammering, as opposed to neurogenic stammering caused by a trauma or stroke (Andy & Bhatnagar, 1992), is a particular type of stammering that has its onset at an early age and lasts throughout adolescence into adulthood. The onset of stammering can be as early as 2 years of age (Costa & Kroll, 2000). Stammering onset by 3.5 years of age accounts for 85% of all instances of developmental stammer with the remaining 15% being spread up to age 12 which is traditionally claimed to be the latest onset for developmental stammering (Howell, 2007). The actual incidence of stammering (i.e., how many people have stammered at any point in time in their life) is as high as 5% (Andrews & Harris, 1964; Brocklehurst, 2013; Månsson, 2000). This accounts for all children who have stammered for a period of at least 6 months. A high rate of natural recovery later reduces the prevalence of stammering (i.e., the number of people who stammer at a given point in time) to approximately 1% (Bloodstein & Ratner, 1969; Craig, Hancock, Tran, Craig, & Peters, 2002). The remaining 1% have a persistent developmental stammer, as their stammer does not recover naturally. These numbers stem from the USA, where a stammer affects roughly 3 million people. It is assumed, however, that these numbers are not country specific, but rather universal (Proctor, Duff, & Yairi, 2002; Yairi & Ambrose, 2005) with worldwide more than 68 million people who stammer.

Research on stammering has targeted multiple areas spanning from genetics and neurology to psychology and linguistics (Büchel & Sommer, 2004; Costa & Kroll, 2000). These approaches appear to be justified seeing that stammering is a multi-dimensional disorder. Research in the field of genetics, for example, has shown that some people may have a genetic predisposition to stammering. Stammering occurs more frequently in people who have relatives who stammer. From twin studies (Howie, 1981), we have learned that monozygotic twins are more prone to both stammer (concordance of 60%) compared to dizygotic twins (concordance of 20-25%). Another genetically linked finding is the increasing gender imbalance with an increasing age of the people who stammer. While at the time of onset both boys and girls appear to be similarly affected by the stammer, a larger natural recovery in girls leaves a noticeably higher percentage of males who stammer. By adulthood, the ratio of male to female people who stammer is roughly three to one.

Neurologists have shown for adults as well as for children that stammering may be related to neural activation abnormalities (Braun et al., 1997; Chang, Kenney, Loucks, & Ludlow, 2009; Rosenfield, 1980). This has been related to a reduction or disruption in white matter tissue (Beal, Gracco, Brettschneider, Kroll, & De Nil, 2012; Chang, Zhu, Choo, & Angstadt, 2015; Connally, Ward, Howell, & Watkins, 2014), which is seen to reflect the neuroanatomical connectivity inhibiting proper control of movements as well as their timing. In addition, differences in cerebral lateralisation were shown to differ between PWS and PNS, where PNS typically display a larger activation in the left hemisphere when processing speech while in PWS the activation is more symmetrical. Directly linked to speech production, a recent study found evidence for reduced activation in AWS in the area 44, a subregion of Broca's area that is responsible for sequencing, motor planning and action inhibition (Neef et al., 2016). In children, however, the effect could not be shown, suggesting the neural differences to stem from neuroplastic adaptation to the stammer (Sowman et al., 2017). Neurologists have further shown the relationship of an imbalance of dopamine levels and stammering in AWS (Alm, 2004; Civier, Bullock, Max, & Guenther, 2009, 2011a; Wu et al., 1997). High levels of

dopamine can be encountered, with dopamine receptor antagonists preventing the absorption of the excessive amount of dopamine and being potentially responsible for the stammer (Civier, Bullock, Max, & Guenther, 2011b; Wu et al., 1997).

Outwith the areas of genetics or neurological research, there have been long-standing theories linking a speakers' psychological wellbeing to stammering (Bloodstein & Ratner, 1969; Johnson, 1930; Starkweather & Gottwald, 1990) . A link could be found between stammering and certain personality traits. Perfectionism or anxiety, for example, have been claimed to stand in the way of fluent speech in people who stammer (Alm, 2014; Brocklehurst, Drake, & Corley, 2006; Iverach & Rapee, 2014; Sheehan & Voas, 1954). This could well blend in with the linguistic stages children go through during language acquisition. It is established that children are re-ordering the way they process speech around the age of two to three years. Children move from holistic to incremental production of words (Anderson, 2007; Oppenheim, Dell, & Schwartz, 2010; A. M. Peters, 1976; Throneburg, Yairi, & Paden, 1994; Wolk, Edwards, & Conture, 1993), which may pose too big of a challenge for some children at that age. With factors of personality and linguistic demand coming together it could well be that some children are overburdened with the shift to more analytical processing of speech. The shift in speech processing could also explain the focused onset of stammering around the ages of 2 to 4 years of age for most cases of developmental stammering.

Common to the various approaches to stammering are the linguistic criteria that are used to establish the two groups of PWS and PNS. Only after speakers are categorised as PWS or PNS, the two speaker groups are compared to investigate potential correlations with respect to the aspect under investigation (e.g., personality traits, dopamine levels). Further studies explored predictors of stammering in CWS such as gender (Reilly et al., 2013). What has not been done by geneticists, psychologists, or neurologists so far, however, is to establish non-linguistic criteria that would hold to classify and possibly even predict pathological stammering. While most of these fields in stammering research could find

correlations with the symptoms of stammering, no clear causality could be established. The linguistic definitions are used to objectively classify a speaker's speech profile as either non-pathological fluent or pathological disfluent. As described by Guitar (1998), this includes frequency, ratio and duration of stammered speech. The frequency count can be performed on a read passage, a monologue or dialogue. Word- and part-word / syllable repetitions, prolongations and blocks are regarded as stammered (Guitar & Belin-Frost, 1998). The syllables stammered are then related to the overall syllables spoken to retrieve a ratio of stammered speech. The duration stammered is often derived from the longest stammer events observed in the speech sample (Riley & Bakker, 2009; Teesson, Packman, & Onslow, 2003).

Speech profiles are considered objective measures. One needs to keep in mind, however, that these profiles are often limited to what can be perceived from the outside. It is evident that much more of the speech signal is available than what can be perceived. Speech judgements are often based on what is perceptually salient (in contrast to what can be measured in the acoustic signal or even articulatory movements that are not captured in the acoustic signal). And where physical concomitants are judged (see The Stuttering Severity Instrument, Riley & Bakker, 2009) these are also limited to what is visually accessible. Hence, accessing both the acoustic and articulatory detail of speech will provide a more holistic picture and potentially help us understand what underlies stammering.

In the next section we will explore the main linguistic theories associating stammering with impairments in language production and speech motor control.

1.4.1 Stammering as a Language Impairment

Stammering is often treated as a language impairment where linguistic complexity plays a major role. The general assumption is that PWS produce typically fluent speech with local breakdowns in fluency where linguistic complexity serves as a trigger to these events of stammering.

1.4.1.1 Levelt's Serial Model

As mentioned earlier, in his model of serial speech production Levelt (Blackmer & Mitton, 1991; Levelt, 1989; Levelt et al., 1999) introduced several stages of speech production. Errors could occur at each of these stages with the nature of the error depending on the stage at which it occurred. Errors at the lexical selection stage would account for word exchanges while errors occurring at the syntactic level would lead to word stranding. Word stranding is when words are incorrectly inserted into slots. Errors in phonological encoding would further account for sound errors.

In addition to the feedforward approach to speech production Levelt incorporated a feedback loop (Levelt, 1983, 1984) whereby speakers can self-monitor their speech. During speech production, the speech plan is investigated and repaired if erroneous. This loop enables speakers to detect and repair speech errors. Depending on the stage where the error is detected and repaired, an error may result in a hesitation or pause and may remain covert (not fully articulated) or if detected at a later stage may become overt. Arenas later investigated the interaction of the speech production and the monitoring systems. He claimed that their interaction would account for the variability observed in stammering (Arenas, 2012). In his thesis, Arenas found a positive correlation between stammering severity, the anticipation of stammering and instances of stammering. Arenas, however, was not the first to establish a relationship between monitoring and stammering as we will see in the following sections.

1.4.1.2 Covert Repair Hypothesis

The Covert Repair Hypothesis (Postma & Kolk, 1993) proposes that breakdowns in the fluency of speech of people who stammer as well as people who do not stammer reflect attempts of the speaker to repair phoneme selection errors during speech production. Depending on the stage at which errors are detected and corrected as well as on the ease with which they can be repaired, disfluencies may

remain covert and unnoticed or they may become overt, i.e., noticeable to the listener.

Postma and Kolk argue that two factors are responsible for (a) the larger number and (b) the more severe nature of the breakdowns in the speech of PWS. PWS make more severe errors due to a less efficient speech production system with both (i) slower activation and (ii) slower repair. Postma and Kolk claim that the slower activation of phonological segment nodes is responsible for more inappropriate nodes to be activated and to compete with the appropriate node (see Dell's spreading activation model; Dell, 1986) causing the phonetic plan in PWS to more frequently become erroneous. In addition to the slower activation, AWS are also slower in repairing the errors which leads to greater and more apparent disruptions in the fluency of speech which are more likely to become overt (Brocklehurst, 2008). ANS compared to AWS are generally successful in noticing and correcting errors early and repairing their errors covertly.

For children who stammer Nippold (2001) shows that the performance is comparable to that of their fluently speaking peers in both phonological encoding and phonetic realisation. During adulthood, however, slower phonological encoding can be observed in AWS compared to ANS (Sasisekaran, De Nil, Smyth, & Johnson, 2006), which could be due to a difference in how PWS and PNS shift from holistic to incremental speech processing (Byrd, Conture, & Ohde, 2007; Melnick, Conture, & Ohde, 2003).

It is established that children compared to adults rely more heavily on holistic processing, which may account for the onset of developmental stammering in childhood and the high rate of natural recovery from stammering. Early language acquisition is holistic in nature. Children learn words as entities and only later move to incremental processing of speech. The peak age for stammering onset coincides with the point at which it is thought children shift from holistic to incremental phonological processing (Byrd et al., 2007). It has been suggested that difficulty achieving smooth transitions between sounds – once speech begins to be produced

incrementally, may be associated with the onset of stammering (Costa & Kroll, 2000; Johnson, 1959; Starkweather, Hirschman, & Tannenbaum, 1976).

This is in line with children showing larger priming effects when primed with rhyme primes. Adults on the other hand process speech incrementally (Kempen, 1987), which is reflected in larger priming effects when primed with target onsets (Brooks & MacWhinney, 2000). In typically developing children, this shift could be shown to occur between the ages of 3 and 5 years. A facilitation effect was shown to increase as a function of age when children were primed with the consonant onset (and the formant transition and most of the vowel of the target). In children who stammer, however, no such effect could be observed leading to the conclusion that CWS are delayed in the shift to incremental speech processing. A delayed and slower shift in CWS could lead to lower proficiency in phonological encoding, which may be reflected in the larger number of phonological errors. The Covert Repair Hypothesis (CRH) accounts for the three main symptoms of stammering, i.e., repetitions, prolongations, blocks.

1.4.1.3 Vicious Circle Hypothesis

Similar to the Covert Repair Hypothesis put forward by Postma and Kolk (Postma & Kolk, 1993), the Vicious Circle Hypothesis (Vasic & Wijnen, 2005) also explains the disfluencies as resulting from attempts to correct errors internally. In contrast to the CRH, however, Vasic and colleagues (Vasic & Wijnen, 2005) do not explain the increased number of disfluencies in the speech of PWS by larger errors that would need repaired. Instead, they suggest that PWS are prone to be overly sensitive to errors. The increased sensitivity in self-monitoring would then account for the detection and attempt to repair even subtle phonetic irregularities (R. J. Lickley, Hartsuiker, Corley, Russell, & Nelson, 2005). The heightened sensitivity might even lead to the inaccurate detection of absolutely perfect productions as having an error in them. Perhaps these more peripheral productions would go unnoticed or not be considered problematic by PNS which would then lead to fewer instances of error repair in PNS.

Vasic and Wijnen further proposed that the overly sensitive monitoring identifies and attempts to repair even disfluencies occurring during error repair resulting in a vicious circle whereby self-monitoring may be further increased by increased anticipation of disfluency (Jackson, Yaruss, Quesal, Terranova, & Whalen, 2015).

1.4.1.4 EXPLAN Hypothesis

The EXPLAN Hypothesis (Howell & Au-Yeung, 2000), in contrast to the Covert Repair Hypothesis and the Vicious Circle Hypothesis (Vasic & Wijnen, 2005), does not see disfluencies as resulting from processes of monitoring and covert error repair. Instead, errors are suggested to directly result from an incomplete speech plan at the time of execution. Blackmer and Mitton (Blackmer & Mitton, 1991) account for disfluencies in typical speech through the incomplete phonetic plan leading to automatic restarts that become acoustically salient.

Howell and Au-Yeung formulate their EXPLAN Hypothesis. The term EXPLAN contains two major stages of speech production. First, there is the linguistic plan (PLAN) that via motor processes is to be executed (EX). Both planning and execution are parallel processes that progress independently. Howell and Au-Yeung (Howell & Au-Yeung, 2000) suggest that the main cause for disfluencies in the speech of PWS is the speech planning that is too slow to keep up with the output speed required for execution. The slow speech planning falls behind the faster speech execution which then runs out of material to be executed. This lag between planning and execution increases with complexity of the linguistic material, which as a result also increases likelihood of stammering. As complexity increases, planning time increases, and the increased planning time may then lead to the speaker running out of material to be executed causing the speaker to stammer. Howell (Howell, 2004) follows the characterisation introduced by Throneburg, Yairi and Paden (1994) who establish three parameters with which complexity can be measured. Their parameters include:

- the type of consonant strings: complexity increases with increasing number of adjacent consonants; /sɪp/ is easier when compared to /strɪp/)

- early vs. late-emerging consonants: complexity increases the later consonants are acquired; /bɒb/ is less complex when compared to /ʤɒb/ because bilabial consonants are acquired at an earlier age (Sander, 1972)

- the number of syllables: the more syllables, the longer the linguistic material and the more complex becomes the coordination of planning and execution of that material.

Howell and Au-Yeung (Howell & Au-Yeung, 2000) propose two surface forms that arise from the situation that the planning is lagging behind speech execution. In their EXPLAN Hypothesis they formulate the concepts of 'stalling' and 'advancing', which to them are two different ways the PWS react to the situation of having run out of linguistic material. 'Stalling' behaviours, like prolongations or blocks, according to Howell and Au-Yeung come about because of delayed planning. The speaker waits for more speech material to be available for execution. In the meantime, he produces a disfluency in the form of a prolongation or block. 'Advancing' behaviours, like repetitions, on the other hand occur when the speaker attempts to execute the speech plan despite the plan being incomplete at the time of execution. In the attempt to advance, the speaker executes repetitive restarts of the already available speech plan which results in repetition-like disfluencies.

1.4.2 Stammering as a Speech Impairment

In the following section, we will explore stammering from the perspective of it being a speech impairment. Disfluencies are perceived as local instances within an otherwise smooth flow of speech. Acoustic and articulatory instruments, however, make it possible to go beyond what is perceived. These instruments make available details about PWS speech that are otherwise not perceptually salient. With this possibility at hand, the question has been raised whether the speech of PWS is indeed typically fluent speech with intermittent local stammer-like disfluencies or whether it is more globally deviant where the stammer affects speech throughout. Support for the latter could imply that higher neurological levels are involved that

affect the speech of PWS more generally (Ludlow & Loucks, 2003). While differences might not necessarily surface to the perceptual level, we expect to observe differences in even the fluent speech of PWS.

1.4.2.1 Fault-Line Hypothesis

The Fault-Line hypothesis (Wingate, 1988) responds to findings that AWS parse phrases based on syllables rather than utterances and that disfluencies typically occur on the first syllable of a word which also carries linguistic stress.

Wingate sees syllables as asymmetric entities. The syllable-initial position is usually occupied by a consonant (syllable onset), which is then followed by a vowel (nucleus) with optional consonants following that vowel. In line with the Coupled Oscillator Model (1.3.3.1.3) he claims that the initial consonant and vowel segments have a relationship different from the relationship between the vowel and the optional consonant following that vowel.

Wingate states that the initial consonant and the subsequent vowel in a CV sequence are produced as "one continuous flow of action […] an intricate blending of the complex muscle systems" (Wingate, 1988, pp. 1982 f.). To support his argument, he refers to studies pointing out the effect coarticulation has on syllable-initial positions where both constituents of the CV sequence, the initial consonant and the subsequent vowel nucleus, are initiated at approximately the same time (Kent & Moll, 1969), also stating that the same is not necessarily true for syllable-final position (MacNeilage & DeClerk, 1969). Difficulty in timing in PWS was also previously suggested by van Riper (van Riper, 1982) who observed productions of schwa indicating a lack of coarticulation in syllabic repetitions. The Fault-Line hypothesis could thereby also account for part-word repetitions like /bə\bə\bɛtə/ where the vowel in the repetition is not fully achieved, but a reduced version of it. With regard to consonant clusters, disfluencies on consonants initial to consonant clusters could constitute difficulty with the consonant itself, but may also be accounted for by anticipation of a problem with the integration of the vowel, which would be in line with the Fault-Line hypothesis (Wingate, 1988).

Further support for the Fault-Line hypothesis is found in studies he conducted himself: Wingate (Wingate, 1982) has investigated stress pattern curves for both speaker groups focusing on the initial peak and the subsequent decay in first syllables. His findings show larger relative amplitude difference between the initial peak and the subsequent decay for AWS when compared to typical speakers. In a follow-up study Wingate (Wingate, 1984b) found evidence for a correspondence of disfluencies and stress peaks which he interprets to account for the occurrence of disfluencies predominantly in syllable-initial position (Wingate, 1969b, 1984a).

The relationship between syllable stress and stammering is also discussed by Howell who investigated the relationship between disfluency, stress and content words as compared to function words for English and Spanish. For the case of English, content words are more often disfluent than function words (Howell, Au-Yeung, & Sackin, 1999). This however does not necessarily hold across all age groups. Younger speakers tend to be more disfluent on function words while AWS are more disfluent on content words (Howell et al., 2004). For the case of Spanish, Howell and colleagues found function words to be more often disfluent – especially in younger speakers. With increasing age, speakers were more often disfluent on content words which is in line with the patterns observed for English speakers (Ardila, Ramos, & Barrocas, 2011; Au-Yeung, Gomez, & Howell, 2003; Dworzynski, Howell, Au-Yeung, & Rommel, 2004).

Wingate claims that the main cause of disfluencies is the physiological difference of consonants and vowels more generally (Wingate, 1988) and the change in phonation when transitioning between consonant and following vowel more specifically (Wingate, 1976). This leads him to hypothesise that PWS do not struggle to initiate the initial consonant (i.e. the syllable onset) or the following vowel (nucleus), but to transition between them (Wingate, 1969b). The Fault-Line hypothesis (Wingate, 1988) therefore postulates that disfluencies result from PWS difficulty transitioning from the syllable onset (typically a consonant) to the nucleus (typically a vowel), which Wingate refers to it as an intra-syllabic event.

In line with more recent brain imaging studies, Wingate proposes an underlying neurological impairment to stammering. Wingate stated that: “although the stammer event finds expression as a breakdown at the level of execution, that is, in motor performance, it seems likely that the fault extends from higher levels of the hierarchy of neural organization for language expression, through several stages of the process that extends from the plane of verbal formulation to the level of final motor execution.” (1988, p. 184). Assuming an underlying neurological impairment, we can expect to find differences more globally in even the fluent-sounding productions of PWS when compared to typical speakers.

1.4.2.2 Directions into Velocities of Articulators Model

The Directions into Velocities of Articulators model (DIVA; Guenther, 1994) is a neural network model of speech acquisition and production. It uses feedback and feedforward control systems in its approach to speech production (see Figure 8).

For speech acquisition, the model suggests that children rely heavily on feedback control as accurate feedforward control is not yet developed. The feedback control system incorporates auditory and somatosensory information, which are represented and interconnected in the different (temporal, frontal and parietal) areas of the brain. Children use auditory feedback and map it to the articulatory information. Auditory feedback is used as a corrective mechanism that also feeds into the forward control model. This narrows the auditory target space for the intended phoneme and develops the feed forward control with each attempt that a certain phoneme is produced. Over time, auditory feedback becomes less important and speakers rely increasingly on feedforward control. With each production, the auditory target regions become smaller and speech clearer. The distance between the target regions increases with decreasing size, which reduces coarticulation.

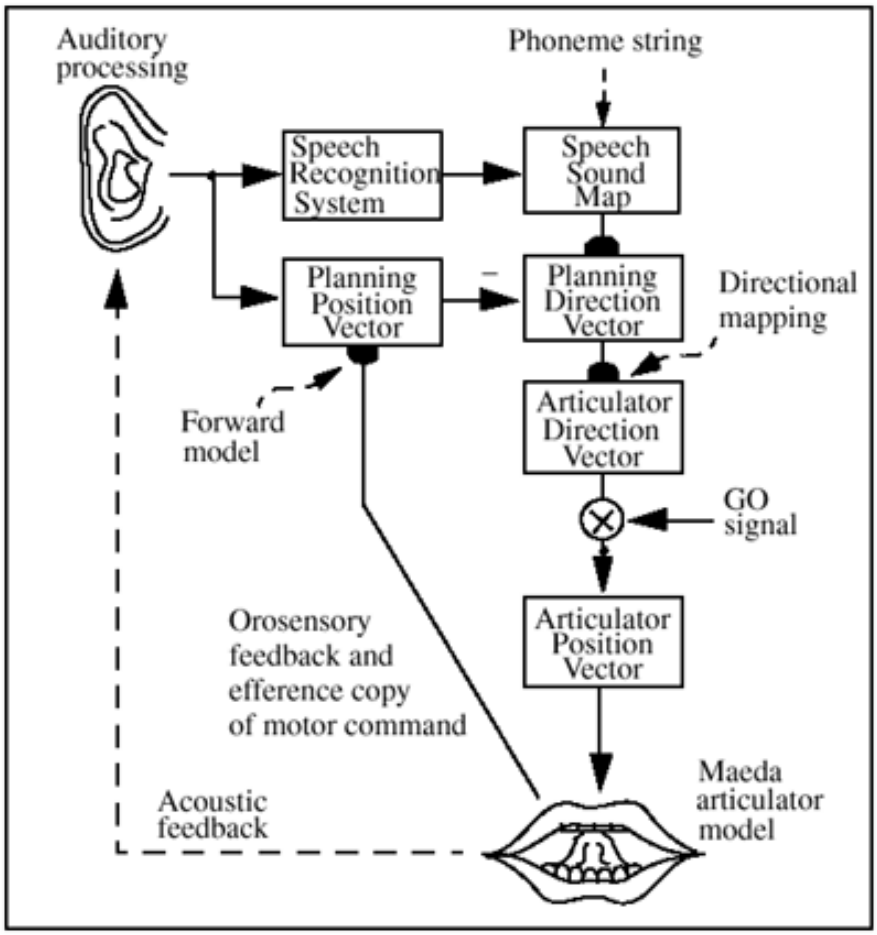


Figure 8 *Schematic representation of the DIVA model (Guenther, 1994, p. 4)*

The DIVA model is important for this thesis as the model has been adapted to speech impairments more generally (Terband, Maassen, Guenther, & Brumberg, 2009) and to stammered speech more particularly (Max, Guenther, Gracco, Ghosh, & Wallace, 2004). Applying the DIVA model to stammered speech, Max and colleagues (Max et al., 2004) proposed two hypotheses: First, that internal models are unstable or insufficiently activated and second, a weaker feedforward or overreliance on feedback control in PWS. The first model proposes that PWS have difficulty developing an accurate and stable mapping between motor commands and the resulting auditory and somatosensory information. The insufficient mapping means that feedforward control cannot accurately predict the sensory consequences, which may cause additional corrections led by feedback

mechanisms. The second hypothesis expects increased difficulty in faster speech as the lag between the motor command and the resulting auditory and somatosensory consequences decreases leaving PWS less time to rely on the feedback and repair.

While the current study does not allow to confirm one or the other, it does allow us to explore a mechanism directly affected by impaired feedforward control mechanisms (Terband et al., 2009), which is anticipatory coarticulation.

1.4.2.3 Monitoring & Coarticulation

For the production of fluent speech, feedforward and feedback control mechanisms need to be in balance. Overreliance on one or the other will most inevitably give rise to disfluencies. For the case of overreliance on feedback, speakers will be more sensitive do errors. Detection of errors means that the production process will be halted while the correction of that error is formulated / encoded. Because the correction is being encoded in isolation, this will affect the degree of anticipatory coarticulation. But even in apparently fluent speech, we can expect PWS to have lowered anticipatory coarticulation as more of the available resources are going into either monitoring or the consequences of monitoring (Postma & Kolk, 1993; Vasic & Wijnen, 2005).

While monitoring affects the capacities available for anticipatory coarticulation, coarticulation in return may also affect monitoring and even trigger disfluency. Coarticulation moves the gesture from the centre of its notional target region. If the target region is defined overly narrow, the coarticulated centre of that gesture may be caught as aberrant and reported back for repair via the feedback loop.

With increasing sensitivity for these coarticulated gestures, not only patterns of stammered productions, but also those of coarticulation would go into error repair, thereby further decreasing the capacities available for anticipatory coarticulation. It might even possible that coarticulation becomes increasingly removed from patterns of stammered productions until the point of release. In addition, as mentioned earlier, coarticulation is a motor skill that is acquired over time. With

monitoring and error repair interrupting the fluent production of speech, the acquisition of anticipatory coarticulation is likely to also be interrupted.

1.4.2.4 Wingate, AP & Coarticulation

In the late 1980s Wingate emphasised physiological aspects of stammered speech and the importance of timing relations between consonants and vowels. In this he pre-empts the concerns of Articulatory Phonology and its implications for coarticulation.

Relating Wingate to Articulatory Phonology using the coupled oscillator framework, we find similarities in how both understand a difference between the CV / in-phase and VC / anti-phase.

Both, however, present two apparently contradictory hypotheses:

1) AP and the oscillator framework predict that in-phase CV syllables are easier in production when compared to VC syllables, which are produced anti-phase. Anti-phase productions are reduced to in-phase productions with increasing challenges.

2) The Fault-Line hypothesis, in contrast, predicts CV syllable production to be more difficult due to the tight coupling of the two gestures for consonant and vowel.

These two apparently contradictory conceptualisations highlight the importance of studying coarticulation and stammering.

A number of studies have explored the speech of people who stammer from an acoustic perspective in AWS (Blomgren, Robb, & Chen, 1998; Dehqan, Yadegari, Blomgren, & Scherer, 2016; Howell, Williams, & Vause, 1987; Maruthy, Feng, & Max, 2017; Prosek, Montgomery, Walden, & Hawkins, 1987) as well as CWS (Chang et al., 2002; Howell & Williams, 1992; Lenoci & Ricci, 2017). Although stammered repetitions of vowels are often perceived as schwa-like, acoustic analysis has revealed that these productions in fact possess the appropriate formant

frequencies for the target vowel but shorter durations (Howell & Williams, 1992). Conversely, there is evidence to suggest that people who stammer have generally greater vowel centralisation than people who do not stammer (Blomgren et al., 1998). This apparent difference in vowel space may, however, relate to organic group differences in speakers' vocal tract dimensions as opposed to functional differences in the speech production process (Prosek et al., 1987).

Evidence concerning the transition from the consonant to the vowel in CV(C) syllables is ambiguous with respect to whether there exist group differences between PWS and PNS. Spectral coefficient measures did not reveal differences between groups (Maruthy et al., 2017). Temporal measures focusing on the second formant (F2) revealed differences in extent and duration but no difference in transition rate between groups (Dehqan et al., 2016).

With regard to velar consonant-vowel coarticulation as measured using ultrasound tongue imaging there was greater variability across children and PWS than the PNS group but no overall group differences in the degree of coarticulation (Frisch & Maxfield, 2017). Using nearest neighbour distance measures on ultrasound data (Zharkova & Hewlett, 2009), children who stammer were found to produce larger Euclidan distance values than children who do not stammer (Lenoci & Ricci, 2017). This suggests that CWS were less stable in even their fluent speech than age-matched control speakers. Lenoci and Ricci (Lenoci & Ricci, 2017) further applied locus equations to the kinematic data obtained using ultrasound. Results showed larger slope values for CWS when compared to CNS, indicating increased variability in CWS when compared to CNS confirming previous findings (Frisch, Maxfield, & Belmont, 2016; MacPherson & Smith, 2013).

1.4.3 Summary and Objective

To conclude these sections on typical speech (section 1.3) and stammered speech (section 1.4), the objective of this study is re-stated: to explore CV transitions in the fluent speech of people who stammer and test the transition deficit proposed by

Wingate (1988) employing measures of timing and coordination. Data produced by PWS will be categorised as perceptually fluent and disfluent. Only data categorised as fluent will be included in the subsequent analysis to explore whether even the fluent speech of PWS deviates from that of typical speech – potentially indicating that stammering does not consist of local events intermittent to typical fluent speech, but that it might affect speech more globally. Materials will consist of CV utterances with varying consonants and vowels as to cover a variety of place and manner of articulation and to discover potential differences in coarticulatory resistance.

The next section will review approaches to measuring motor control more generally and in the speech of people who stammer more particularly. In addition, we will consider the utility of ultrasound tongue imaging to measure lingual kinematics in people who stammer.

1.5 Measuring Motor Control

Under experimental conditions, PWS perform more poorly across a range of acoustic measures of speech performance when compared to PNS. PWS often perform with slower and more variable motor performance when compared to PNS (H. F. M. Peters, Hulstijn, & van Lieshout, 2000 for review; van Lieshout, 1995).[3]

1.5.1 Acoustic Measures of Timing

Based on the acoustic signal alone, PWS have been found to perform poorer across multiple measures. PWS exhibit longer and more variable speech reaction times (Cross & Luper, 1979; Harbison, Porter, & Tobey, 1989; Horii, 1984). Group differences between PWS and PNS in voice onset times (VOT) may be observable only in specific phonetic or utterance contexts (De Nil & Brutten, 1991; Healey & Ramig, 1986; Watson & Alfonso, 1982). When compared to PNS, PWS as a group

[3] Parts of this chapter have been previously reported in Heyde et al. (Heyde, Scobbie, et al., 2016).

have been found to have longer vowel and consonant durations (M. R. Adams, 1987; Di Simoni, 1974; Starkweather & Myers, 1979). PWS were found to have descriptively longer closure durations (Borden, Kim, & Spiegler, 1987) and VOT than PNS (M. R. Adams, 1987; Bakker & Brutten, 1990).

Slower motor performance is often interpreted as an indicator of motor deficits (Zimmermann, Smith, & Hanley, 1981). A different interpretation sees these delays in timing as resulting from compensation strategies. Slowing the execution of movements decreases the demands, making them more manageable for AWS (van Lieshout, Peters, Starkweather, & Hulstijn, 1993). Further studies found evidence for a strong relationship of segment durations, which could not be shown for CWS (Zebrowski, Conture, & Cudahy, 1985). But not only slower, also more variable performance is interpreted as deficits on the motor level as could be shown for both AWS and CWS (Dokoza, Hedever, & Sarić, 2011; Frisch et al., 2016; Jäncke, 1994; Onslow, Van Doorn, & Newman, 1992; Packman, Onslow, Richard, & Van Doorn, 1996; Perkell & Klatt, 2014; van Lieshout, Namasivayam, & Maassen, 2010).

A different approach to motor performance is the investigation of not just the segments, but of the transitions between segments as implied by Wingate (Wingate, 1969b, 1988). In the following section, we will discuss on the examples of locus equations and formant slopes what formant frequencies can reveal about the coordination of transitions in speech.

1.5.2 Acoustic Measures of Coordination

Formant frequencies are understood as indicators of articulatory (pharyngeal and oral cavity) space. The position of the articulators changes the oral cavity size, which then changes the preferred resonating frequencies (referred to as formants F1, F2, and F3) on the frequency response curve (Ladefoged, 2006).

Tongue body position has a main effect on F1 and F2 reflecting the vertical (height) and horizontal (frontedness) dimensions of the tongue body. The central vowel /ə/

is produced centrally at approximately F1: 500Hz, F2: 1500Hz. Moving the tongue body on the vertical plane (raising or lowering) will affect the pharyngeal space that can be observed in the first formant, F1: the higher the tongue body, the more pharyngeal space, the lower the F1 frequency and vice versa: the lower the tongue, the less pharyngeal space, the higher the F1 frequency. Raising to produce /i/, for example, will change F1 to ~300 Hz while lowering for /ɑ/ increases the formant value for F1 to ~800 Hz.

Moving the tongue body on the horizontal plane (fronting or retracting) affects the oral cavity correlating with the second formant, F2: the more retracted the tongue body, the larger the oral cavity, the lower the frequency and vice versa: the more advanced the tongue body, the smaller the oral cavity, the higher the F2 frequency. Tongue frontedness / retraction will therefore change F2 to ~2100Hz for fronted vowels like /i/ and change F2 to ~1200 Hz for vowels that are produced with a more retracted tongue body like /ɑ/.

Consonants are produced by creating friction (in the case of fricatives) or full constrictions (in the case of stop consonants) in the oral cavity. Like vowels, frequencies at the burst of the consonant also differ as a function of vocal tract size during the constriction. For alveolar consonants, the vocal tract length in front of the constriction is small resulting in higher F2 frequencies (F2~2500-4000Hz) while the vocal tract length in front of velar constrictions is larger resulting in lower F2 frequencies (F2~1500-2500) (Reetz & Jongman, 2011).

If segments were produced independently, the frequencies at which each segment is produced would remain stable. Consonants and subsequent vowels would each be produced at their own inherent and independent frequencies. In running speech, however, segments are stringed together, which causes neighbouring segments to influence each other as tongue configurations of neighbouring segments overlap.

In the case of anticipatory coarticulation in a CV syllable, the tongue configuration required for the vowel is initiated almost simultaneously with the preceding

consonant. Speakers change the articulation of the consonant in anticipation of the subsequent vowel, which then changes the frequencies for that consonant into the direction of those of the following vowel.

Formant frequencies can help us understand how speakers manage the transition from one speech segment to another. For the case of stammering as a condition, coarticulation could affect transitions in two ways: Co-production of neighbouring segments reduces formant distances of these segments, which may reduce the articulatory effort. In this view, coarticulation is seen as facilitative as it is typically accompanied with formant undershoot (Krull, 1989b) which is considered articulatorily more economic (Kent, 1983; Sereno, Baum, Marean, & Lieberman, 1987). PWS could be expected to perform with a larger overall coarticulatory degree when compared to PNS.

On the other hand, larger degrees of coarticulation also imply a larger overlap of entities that are potentially highly different in their physiology. Mastering the transition in addition to the two articulatory targets preceding and following the transition may add to the complexity and possibly partially account for the breakdowns in stammered CV transitions. To reduce the complexity and allow time for more refined articulatory gestures (Lindblom, 1983; Nittrouer, Studdert-Kennedy, & Mcgowan, 1989), PWS might produce transitions with reduced degrees of coarticulation (Barbier, Perrier, Ménard, Tiede, et al., 2013; Frisch et al., 2016; Zharkova et al., 2011) to maintain fluency.

The following section presents a brief discussion of Locus Equations as one means to explore coarticulation in typical and stammered speech. We will subsequently discuss formant slopes and how they can be of avail to address transitions and the Fault-Line hypothesis more directly.

1.5.2.1 Locus Equations

Lindblom (Björn Lindblom, 1963) was one of the first to introduce locus equations to speech research as a rather general measure of contrast, which were later

applied to CV syllables to look at anticipatory coarticulation (Krull, 1988; Lindblom & Sussman, 2012). Locus equations have been interpreted as an indicator of "the degree of coarticulation between a consonant and a following sonorant" (Reetz & Jongman, 2011). This measure investigates the variation of F2 values at the onset (i.e., first glottal pulse after the release burst) of a consonant together with the F2 values at the target (i.e., midpoint) of vowels following that consonant.

Locus equations have been shown to be a valid estimate of the degree of consonant and vowel coarticulation (Geitz, 1998; Sussman, Hoemeke, & McCaffrey, 1992). The more independent a consonant is from the following vowel, the more consistent is the F2 value of that consonant across vowel environments. On the other hand, the larger the variation in F2 at the onset of the consonant in response to the following vowel, the larger the degree of anticipatory coarticulation.

Robb and Blomgren understand coarticulation as an opposing force to fully refined articulatory targets. Robb and Blomgren (1997) found flatter F2 slopes suggesting less gestural overlap, i.e., lower degree of coarticulation between phonetic segments. According to them, the lower degree of coarticulation in PWS allows for a more refined individual gesture, which also fits in with the previously mentioned slowing of movements to decrease demands and make motor commands more manageable (van Lieshout et al., 1993).

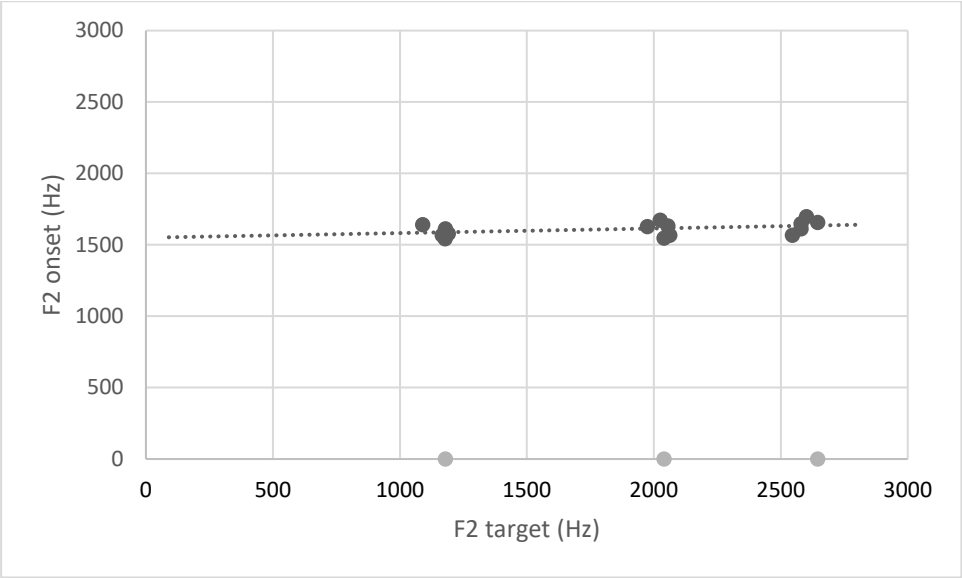


Figure 9 *A simulated example of a regression fit to F2 values obtained at the consonant onset produced in three vowel environments: The very flat regression indicates very little to no coarticulation.*

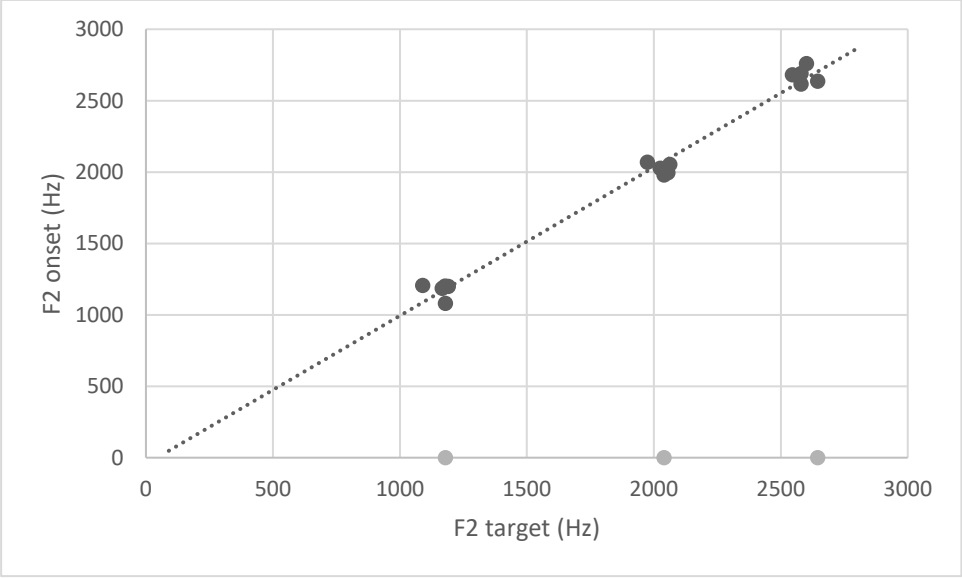


Figure 10 *A simulated example of a regression fit to F2 values obtained at the consonant onset produced in three vowel environments: The very steep regression indicates large degrees of coarticulation.*

Typically, a larger sample of recordings is employed to get an impression of coarticulatory degree for a consonant in different vowel environments by speaker or by speaker group. To measure and quantify the degree with which F2 at the onset of a consonant varies, linear regressions of F2 onset at CV transition and the target frequency at the vowel are calculated (Sussman, Hoemeke, & Ahmed, 1993, p. 1256). F2 values are extracted for each production of a V_PCV utterance. Locus equations then use the F2 value at the onset of the CV transition for a consonant and relate these to the F2 value at the target of different vowels. The slope of the resulting regression line then indicates the degree of coarticulation for that consonant.

The steeper the regression fit, the more variation can be observed in the F2 values of a consonant as a function of the following vowel, the larger the degree of anticipatory coarticulation (see *Figure 10*). In contrast, if there was no coarticulation from a particular consonant to the different vowels, the regression line would be a near flat line with a slope approaching zero (see *Figure 9*). Hence, locus equations have found frequent application in the research of coarticulation related to speech production as well as language acquisition (Fowler, 2000; Iskarous, Fowler, & Whalen, 2010; Montgomery, Reed, Crass, Hubbard, & Stith, 2014; Rubertus, Abakarova, Tiede, Ries, & Noiray, 2016; Sussman et al., 1993).

For the purpose of this current study, we will employ Locus Equations to investigate whether overall differences in coarticulatory degree can be detected for PWS – something not predicted by Wingate. Locus Equations are typically used to investigate coarticulation. They inform about what is happening in the transition based on secondary acoustic information. F2 measures are obtained in the consonant state to see the effect the subsequent vowel has on that consonant. An impression of overall coarticulatory degree is obtained by consonant and speaker / group (see Figure 9 and Figure 10).

In his Fault-Line hypothesis, Wingate states that stammering results from a transition deficit with difficulty in the transition, but no difficulty in reaching the

targets to either side of the transition. Because calculations of Locus Equations are based on measures taken at the acoustic target, no direct information can be obtained about potential differences that occur in the transition between targets. As Wingate states, PWS are not expected to perform differently on a consonant or vowel state. According to this hypothesis, no difference for PWS should be found using Locus Equations – an assumption confirmed by previous studies: In a study by Sussman, Byrd and Guitar (Sussman, Byrd, & Guitar, 2011) no significant group difference could be found. PWS locus equations were found to fall within the range of LE of typical speakers. When comparing mild and moderate degrees of stammer severity to that of severe stammers, a tendency for tighter distribution of output with decreasing stammering severity was observed.

1.5.2.2 Formant Slope

Formant slopes are established measures that have been employed to inform about coarticulation through measures of formant slope duration, formant slope extent as well as formant slope transition rate. Formant slopes can provide a more complete picture of transitions when compared to the measures obtained from locus equations. While locus equations are based on two measures (F2 at consonant onset and F2 at the vowel target), formant slopes are based on F2 values that are obtained at the consonant onset and at specified intervals from the consonant onset thereby capturing the entire transition for each recording.

The beginning and end of the transition (F2 onset and F2 target) are established for each recording. For the definition of both F2 onset and F2 target different approaches are described in the literature. F2 onset is broadly defined as the formant frequency at CV boundary for which different measures have been applied. Formant frequencies for F2 onset are measured at the first glottal pulse at the burst release (Chang et al., 2002; Robb & Blomgren, 1997), the first pulse of the vowel (Krull, 1989b; Sussman & Shore, 1996) or initiation of high-frequency spectral energy for fricatives (Robb & Blomgren, 1997). For F2 target frequencies, measures differ equally. F2 target frequencies are obtained at the vowel steady state (Chang

et al., 2002; Sussman & Shore, 1996), a maximum / minimum turning point (Krull, 1987) or fixed time points (Nearey & Shammass, 1987; Robb & Blomgren, 1997) as opposed to visual inspection (Yaruss & Conture, 1993; Zebrowski et al., 1985).

Based on F2 onset and F2 target, three measures are typically obtained: The temporal lag between beginning and end is measured to obtain the transition duration. The difference in frequency between beginning and end of the transition is used to obtain a measure of slope extent. Resulting from these two measures, the transition rate is obtained, which is the ratio of slope duration and slope extent. These measures can be directly related to coarticulation where differences in coarticulation will reflect in differences for the formant slope duration and the formant slope extent. The larger the degree of coarticulation, the more are the formants of a consonant shifted in the direction of the formants of the following vowel. The adaptation to formant values to those of the subsequent vowel reduces the formant difference between consonant and vowel, which is equivalent to a decrease in formant slope extent. When transitioning between consonant and vowel, a reduced formant slope requires a lower articulatory effort also decreasing the formant slope duration.

For research on stammering, formant slopes are a valuable resource as they can be obtained for each recording also providing a more fine-grained picture when compared to locus equations. This is particularly helpful for the current study as we will apply formant slopes to the perceptually fluent speech of PWS and PNS where difference might only be found in the detail.

Temporal and spatial measures might aid in the understanding of differences between PWS and PNS. Additionally, the degree of stability with which these transitions are executed may suggest differences in motor control. Using the coupled oscillator model, van Lieshout (van Lieshout, 2017) has shown that smaller movement amplitudes reduce the coupling stability of articulatory gestures, which might also be reflected in potential differences in variability / homogeneity.

Employing the measures obtained from formant slopes, F2 formant structures were often found atypical in people who stammer (Chang et al., 2002; Robb & Blomgren, 1997; Stromsta, 1986; Subramanian, Yairi, & Amir, 2003; Yaruss & Conture, 1993). Yaruss, Conture (Yaruss & Conture, 1993) compared the stammered and fluent productions of children who stammer and found a positive correlation between stammered and fluent F2 transitions for transition extent and transition rate. A similar observation was made in a study by Subramanian, Yairi and Amir (Subramanian et al., 2003). They investigated F2 frequency change and F2 transition durations in CV transitions for children who do and do not stammer. They found children who do not stammer to exhibit smaller frequency changes when compared to children who do stammer. Chang, Ohde and Conture (Chang et al., 2002) found differences in formant transition rate as a function of place of articulation. Again, children who stammer showed smaller contrasts of formant transition rate between labial and alveolar consonants when compared to children who do not stammer suggesting that the articulatory settings for different places of articulation are not as contrastive and refined in PWS when compared to PNS. In a more recent study, Arnold (2015) used F1 and F2 formant transitions to investigate the overreliance of PWS on sensory feedback. Though no significant group effect was found for neither formant transition duration nor transition rates, trends are reported that are consistent with previous literature where PWS exhibited a trend for longer F1 and F2 transition duration when compared to PNS. Second, in the casual speech condition, PWS performed at lower F1 and F2 transition rates when compared to PNS. Dehqan and colleagues (Dehqan et al., 2016) reported greater mean formant extent for PWS when compared to typical speakers. They found differences in duration where PWS produced longer transition durations, which may contribute to the overall slower syllable rate when compared to PNS.

Though the findings are not all consistent, the studies presented above show the need to evaluate F2 formant structures to assess Wingate's (Wingate, 1964, 1969b) claim that a transition deficit is central to stammering.

1.5.3 Kinematic Measures

The study of tongue movements has advanced tremendously over time. Using the current state of the art, highly precise instruments are available which allow us to access speech most directly at the articulatory level. Kinematics is important but has been under-represented in the literature due to the complexity and inaccessibility of techniques in distinction to the relative ease of acoustic analysis. Articulatory techniques are beneficial as they enable us to study independent articulators directly, which is invaluable for the study of particularly speech motor control.

Optical tracking systems can be used to investigate the movement patterns of lips and jaw with high precision (Feng & Max, 2014; Munhall & Vatikiotis-Bateson, 1998; A. Smith, 1992; A. Smith, Goffman, Sasisekaran, & Weber-Fox, 2012; Walsh & Smith, 2002). These techniques are non-invasive as they use infrared light emitting diodes (IRED) that are attached to flesh points to be tracked. While optical tracking systems are ideal for tracking extra-oral structures, they are also limited to those. For the investigation of the primary articulator supplementary instrumentation is required.

In the following section we will introduce three main methods that have been established in the research of the primary articulator, the tongue (see Babatsouli, 2015 for review). We will present electropalatography (EPG) and electromagnetic articulography (EMA) before turning to ultrasound tongue imaging (UTI), which we employed for the current study.

1.5.3.1 Electropalatography

Electropalatography (EPG) makes lingual contact points at the palate observable. An artificial palate is produced for every participant. The palate contains electrodes that are spread equally from front to back and left to right. Each electrode registers individually whether or not there is lingual contact at a certain point in time. Information from the electrodes together inform about tongue shapes the tongue produces in palatal contact. Relative to a produced sound certain shapes are expected. For /k/, for example, the electrodes in the velar area are expected to

indicate tongue contact with a full closure prior to release. Fricatives such as /s/ on the other hand are expected to show lateral lingual contact due to the airstream travelling through centrally. Comparing expected lingual contact patterns with those produced by clients is used not only for diagnosis, but also for intervention (Carter & Edwards, 2004; Gibbon et al., 2001; Gibbon & Wood, 2010; Howard & Varley, 1995; McCann, Timmins, Wood, Hardcastle, & Wishart, 2009; Öller Darelid, Hartelius, & Lohmander, 2016; Wishart, Timmins, McCann, Hardcastle, & Wood, 2008; Wood, Timmins, Wishart, Hardcastle, & Cleland, 2019; Wrench, Gibbon, McNeill, & Wood, 2002). The advantage of this method is that palates are produced individually, and electrodes are spread evenly across the space available. The space between electrodes is normalised allowing for intra- as well as inter-speaker comparison.

There are two aspect that limit the application of electropalatography: First, there is the financial limitation. EPG requires individually manufactured EPG palates which means that they cannot be mass-produced and are therefore relatively costly. For speech and language intervention with younger clients more than one palate may be required because their palates grow and change in shape while the EPG palates cannot be adjusted. For research purposes, the high cost involved often limits EPG studies to case studies or studies with smaller cohorts regarding the data collection.

Another limitation constitutes the type of data that can be collected using EPG. EPG is only applicable when looking at consonants with lingual palatal contact. Any sounds without palatal contact cannot be investigated using EPG, which excludes labial, labio-dental, and dental consonants as well as vowels more generally. Additionally, consonants produced in pharyngeal places of articulation are excluded. Further, EPG is typically used to investigate lingual patterns statically, while other instruments (see for example 1.5.3.2) are preferred to dynamically investigate lingual movement patterns.

1.5.3.2 Electromagnetic Articulography

Electromagnetic articulography (EMA) is another device employed researching articulation (Ananthakrishnan & Engwall, 2011; Mooshammer, Hoole, & Kühnert, 1995; Murdoch, Theodoros, Stokes, & Goozée, 2000; Perkell et al., 1992; Schönle et al., 1987; Steiner, Richmond, & Ouni, 2013; van Lieshout & Moussa, 2000; Wenig & Conrad, 1987). This technology involves placing electromagnetic coils on the speakers' active and passive articulators, which can then be traced in a three-dimensional space. Typically, three sensors are glued to the speakers' tongue, two to the lips and additional sensors are attached to the chin, and nose for reference. EMA offers maximal temporal resolution with only little error in precision. In contrast to EPG it is not limited regarding place or manner of articulation. Tongue movement information can be obtained during any sound produced. Importantly, this methodology also allows investigation of transition between sounds, i.e., between consonants and vowels. Measures of duration and peak velocity can be obtained and quantified to investigate kinematics for trajectories (Geng et al., 2013; Hoole & Harrington, 2013; Hoole & Nguyen, 1997; Recasens, 2002; van Lieshout & Moussa, 2000; Ward, 1997).

This technology is, however, limited to only few measurement coils that are placed on the speakers' tongue surface thereby reducing the data that could be obtained for the entire tongue surface to only these three midsagittal measurement points. When placing the coils, the common procedure is to place the anterior sensor about 1 cm away from the tongue tip while the most posterior sensor is placed as far back as possible, i.e., where the speaker can still tolerate it. The third sensor is then placed medial to the other two sensors.

Limitations of EMA concern the invasive nature of the setup, the selective nature of the data as well as the high costs resulting from the technology and the labour required. It is categorised as an invasive procedure because sensors need to be glued to the tongue, lips and chin to obtain data. Typically, the tongue takes two to three sensors. The placing of the coils often follows a standard procedure where the

most anterior coil is placed 1cm away from the tongue tip and the most posterior sensor as far back as the speaker can tolerate (Hoole & Nguyen, 1997; Kühnert et al., 2006). Additionally, speakers differ greatly in their anatomy. Tongue sizes and oral cavities are of different size and different shape. Differences in the speakers' anatomy (size and shape of tongue and oral cavity) as well as the individual's tolerance of the posterior sensor make it difficult to normalise the data. While EMA data indicates the movement direction and speed of the point at the tongue where the coil is attached with high precision, it is limited to the three data points of the respective coils.

EMA recordings require a large amount of highly specialised technical knowledge during and after the recording session. Sensors that are glued to the tongue need to be attached midsagittal. Only slight rotations can make any data obtained defective. Sensors are likely to be defective or come off before the recording session has finished, both of which means a great danger of loss of data – especially since the entire tongue surface has already been reduced to only these three sensors capturing the movement data only locally. Combining the tongue and lip movement data with the reference points to align them in time and space requires again highly specialist knowledge for the data post-processing.

Together, the highly specialised technology as well as the labour required result in considerably higher costs when compared to ultrasound tongue imaging. The high costs involved only rarely allow for larger cohorts to be recorded for research – not to mention clinical applications of EMA.

1.5.3.3 Ultrasound Tongue Imaging

Ultrasound Tongue Imaging (UTI) involves an ultrasound probe that is held below the speakers' chin where it sends ultrasound waves out. The signal is fan-shaped and sent midsagittally between the lower jaw bone and the Hyoid bone. The signal travels through the flesh and as soon as it meets a different density, an echo is sent back to the probe. It reaches the biggest difference in density at the tongue surface, which appears as a white line on the ultrasound image. Ultrasound captures

midsagittally almost the entire tongue surface, which stands in stark contrast to EPG and EMA. Depending on the anatomy of the speaker (i.e., distance between hyoid and chin bone) the tongue root and in some cases even the tongue tip can be imaged. Though the image quality is better when lower tongue contours are imaged it is generally possible to image the tongue in palatal contact. The recorded quality depends on the distance the signal is set to travel and the resolution. Depending on the imaging participant both need to be balanced, i.e., larger distance with lower resolution or vice versa. There is always a trade-off between these two factors.

Ultrasound, like EMA, captures kinematic information about the key active oral articulator, namely the tongue. Another aspect that sets UTI and EMA apart from studies that investigate only the external articulators such as lips and jaw is that the tongue is crucial for most consonants and all vowels. But even though the tongue plays a role in consonants and vowels alike, the sequencing and overlap in time and space of different parts of the tongue need to be considered. UTI and EMA are not identical, however, in their suitability for providing such data. When measuring the kinematics of the tongue, EMA typically offers a better temporal and 2D spatial resolution than UTI. There are two aspects, however, where UTI is advantageous over EMA, namely that it provides holistic mid-sagittal tongue surface data, and that its output is not limited to just three or four anterior data points. Also, UTI is more accessible. In terms of spatial resolution, UTI is equivalent to EMA in radial directions relative to the probe (sub millimetre accuracy), but is worse in circumferential measures, both as distance from the probe increases, and as the number of echo pulse beams within a given field of view decreases (Wrench & Scobbie, 2011). Both techniques are poor at imaging the tongue tip, since EMA's coils interfere with articulation, while UTI loses its capacity to image the tip if it is masked by the jaw shadow or raised to create a sublingual air pocket.

Regarding the nature of the kinematic measures, both instruments draw on different underlying spatiotemporal data. While UTI provides images of almost the entire tongue surface moving in time and space in a two-dimensional plane, EMA

tracks the path of a few pre-determined fleshpoints, typically but not necessarily in just two dimensions and just in the mid-sagittal plane. Typically, for EMA, three or four electromagnetic coils are glued on the anterior part of the tongue's upper surface as close to a mid-sagittal site based on the tongue's symmetrical morphology as possible, and nowadays coils are recorded as they move in 3D, with analysis based on a data reduction to 2D movement within a cranial mid-sagittal plane. Ultrasound instead samples movement of the tongue's surface through a single plane and is typically orientated to cranial mid-sagittal orientation. It therefore captures an apparent mid-sagittal image of the tongue from near the tip right down to the root through space and time. This provides information not only about the tongue upper surface shape and location, but about tongue internal muscles (e.g. genioglossus), which can contribute to a principal components analysis. It is still regarded sufficient in most research to consider only the wealth of surface data which both techniques provide, in apparent 2D motion, while remembering the different nature of these idealisations. Since the tongue's midline and the cranial midline need not correspond exactly at rest, and since they vary during speech thanks to slight lateral asymmetries in speech production, the 2D data provided differ at source, even before we approach the holistic vs. fleshpoint differences. Finally, of course, other crucial lateral and constrictional aspects of spatiotemporal production ought to be considered for a full picture, which requires using other techniques, such as Electropalatography or MRI.

Compared to the fleshpoint data from EMA, UTI is particularly relevant for exploratory as well as clinical research as it offers a more holistic image of the tongue (Bressmann et al., 2010; Cleland et al., 2019; Cleland, Scobbie, Heyde, Roxburgh, & Wrench, 2017; Davidson, 2005; Grice et al., 2014; Iskarous, 2013; Lawson, Scobbie, & Stuart-Smith, 2013; Wrench, Cleland, & Scobbie, 2011; Zharkova et al., 2011). This is useful in cases, where we cannot know a priori where exactly to measure kinematics, for example, where the right place would be to place an EMA coil. The place of consonantal constriction may, for example, be more variable for experimental speakers with a speech disorder than for control speakers, and

movement patterns of a coil in a suitable place for typical speech might be unrevealing for disordered speech. It is often not highlighted, in fact, that even for quantifying typical speech, the placing of an EMA coil is crucial, since slightly different coil placement provides a different kinematic trace, and different analytic values. Greater study of how variation in EMA coil placement affects kinematic measures is needed to ensure the validity of data and derived measures. The same is of course true of kinematic measures from ultrasound, as we will see.

UTI is more easily accessible and non-invasive when compared to EMA. This aspect is particularly relevant when recruiting and testing clinical populations of relatively low incidence (for example, at approximately 1% for stammering (Craig et al., 2002)), as UTI can be undertaken by a wider range of research teams and disciplines. The relatively non-invasive nature of UTI (Cleland, Scobbie, Roxburgh, & Heyde, 2016; Wiethan, Ceron, Marchetti, Giacchini, & Mota, 2015; Zharkova et al., 2011) is valuable when working with populations who may be particularly sensitive to and atypical in their adaptations to alterations in sensorimotor feedback, since EMA requires that people speak with wires emerging from between the lips. With UTI most speakers tolerate the headset needed to stabilise the UTI probe. The great advantage of EMA, however, is that the data from each coil is perfectly suited for dynamic analysis, and there is large literature of established techniques (Hoole & Nguyen, 1997; Schönle et al., 1987). On the other hand, quantitative analysis of UTI is typically static (Davidson, 2007; Frisch & Wodzinski, 2016; Zharkova & Hewlett, 2009; Zharkova, Hewlett, & Hardcastle, 2008), in terms of the shape of the tongue at a segmental target. It usually relates the ultrasound data to associated acoustic events relative to which singular ultrasound frames are extracted (whether the acoustic midpoint, stop burst or maximum constriction). Static UTI analysis has been employed to explore articulation from a variety of angles. Video-based ultrasound with data output rates of 30 frames per second (which can be deinterlaced to 60 fps if appropriate) can, however, also be used for timing analysis. For many purposes video rate output is as useful as high-speed ultrasound (Wrench & Scobbie, 2008) and it has been used to investigate socio-phonetic processes of timing (Lawson,

Stuart-Smith, & Scobbie, 2014) and also processes of motor control (Zharkova et al., 2014) including more specifically coarticulation and inter-gestural timing (Gick & Campbell, 2003). Both experimental and theoretical evidence indicate that to investigate stammering it is valuable to explore temporal as well as spatial aspects of speech execution. Despite the optimism of Wrench and Scobbie (Wrench & Scobbie, 2008), the low number of frames used in video ultrasound probably limits such kinematic analyses too much. Not only may the few frames that are available not be able to meaningfully capture the subtle nature of articulatory movement, but more importantly, a slower scan rate at the probe combined with buffering of data to create the images may result in both temporal smearing of the raw image, double tongues, and other spatial artefacts in the output images (Wrench & Scobbie, 2006). These make video data more suitable for analysis of the slow-moving end points of articulatory-acoustic targets (i.e. the targets) than for kinematic analysis, especially of fast-moving articulations (Wrench & Scobbie, 2011). This is particularly important when investigating a disorder which essentially involves disruption to the smooth gestural flow of spoken output, where it is the process of articulatory-acoustic target attainment which is our primary interest.

1.5.3.3.1 Dynamic Analysis

The standard approach to UTI has almost exclusively been static where individual or smaller numbers of static images of tongue shapes were compared. Lower numbers of ultrasound frames are typically analysed statically (Belmont, 2015; Frisch & Wodzinski, 2016). Over multiple repetitions of the same target, tongue contours were extracted, and overlaid. The variation of the overlaid tongue contours is interpreted as a measure of articulatory stability of a target. When extracted at different targets, means of tongue contours can also be used to compare the realisation of those targets (Zharkova, 2016). This approach is typically preferred because the smaller number of frames is more readily interpretable.

For the investigation of transitions, however, it is necessary to understand the kinematics between targets, which requires dynamic analysis. Kinematic

information can contribute meaningful additional information to the study of stammering, particularly in light of the suggestion that stammering is best understood as involving disruption to the high temporal coordination of oral (articulatory) and laryngeal (phonatory) movements (D. C. Adams, 1999; Max & Gracco, 2005; van Riper, 1982; Wingate, 1969b).

Dynamic analysis of ultrasound data, like that of EMA data, needs to be based on a larger number of frames. With the latest technological developments, the raw high-speed ultrasound data captured and stored at 120 frames per second or higher (Wrench & Scobbie, 2011) offers a much higher framerate. With high-speed ultrasound it is possible to capture kinematics of tongue movements in transition. The large number of frames allows in-depth temporal and spatial investigation in principle. Articulatory events can be observed throughout the entire recording enabling the researcher to explore events that are less predictable and not necessarily perceptually or even acoustically salient (Iskarous, 2005b). Both aspects about temporal and spatial resolution of UTI are beneficial for detailed analysis of speech movements, even in qualitative analysis (Scobbie, Punnoose, & Khattab, 2013).

Like EMA, measures of duration and velocity can be obtained, to shed light on the trajectory of the tongue surface and its components. This has been useful in the investigation of degree of coarticulation (Zharkova et al., 2014) and inter-gestural timing movement (Strycharczuk & Scobbie, 2015).

1.5.3.3.2 Quantifying Measures

In order to compare and reliably quantify data within and across speakers, referents for data orientation are required. Previously, attempts were made to define external referents that would allow for data orientation (Zharkova, 2013) based on, for example, the real space or vocal tract size. External referents like vocal tract size and real space, however, differ within and / or between speaker. Vocal tracts differ between individuals (Belmont, 2015; Fant, 1966; Fitch & Giedd, 1999; Kent & Moll,

1969). And referents like ‘horizontal’ and ‘vertical’ that are orientated in real space, change with the speaker’s change in posture. Referents that are extrinsic to the data thereby introduce artefacts that should be accounted for in the analysis. Using a data-intrinsic referent circumvents the problematic of orientation and renders measures comparable within and between speakers.

Tasko and Westbury (Tasko & Westbury, 2002) introduced movement strokes as a measure of articulatory movement that can be quantified. For consonants, intrinsic muscles located within the tongue need to be activated. For obstruents, more particularly, the activation of lingual musculature can be described as ballistic showing a single-peaked velocity profile (Gracco & Abbs, 1986). The ballistic movement includes initial acceleration towards a target and subsequent deceleration to reach the target (Ludlow & Loucks, 2003; Prasad, Kellokumpu, & Davis, 2006; Shawker, Sonies, & Stone, 1984). The bell-shaped ballistic movement describes an interval with a local peak in speed. These ballistic movements can be broken down into movement strokes, which are periods of acceleration and deceleration each with a peak in velocity. These movement strokes are determined using data-intrinsic kinematic features such as peak velocities and troughs (Tasko & Westbury, 2002). Movement strokes lend themselves to the study of stammered speech for three reasons:

- bell-shaped movements are considered skilled movements, which makes them a good use case for the study of motor impairments in comparison to typical speakers
- the resulting measures, i.e., movement strokes are defined through data-intrinsic referents, which makes them independent from any type of speech unit
- strokes can be identified automatically, which means that the method with which these measures are derived is consistent and therefore replicable

While these movement strokes can be applied to any type of motor task, they are not suited for all kinds of data. The limitation of these movement strokes is that they are most suited for point-tracking data such as electromagnetic articulography. In this thesis, we will offer a solution that utilises movement strokes to quantify UTI data.

1.5.4 Summary and Objective

To conclude this section on measures of motor control, the objective of this thesis is re-stated: to apply a new method using ultrasound tongue imaging to obtain measures of tongue kinematics in perceptually fluent speech of people who stammer. This thesis purely focuses on CV transitions. Established acoustic measures of timing (segment durations) and coordination (formant slopes and locus equations) will be applied to CV utterances in the perceptually fluent speech of PWS and the speech of control speakers. Findings will be assessed against findings of previous studies. Lingual kinematics from CV utterances in the fluent speech of PWS will be explored using ultrasound tongue imaging to investigate the transition deficit hypothesis. Finally, kinematic findings will be discussed against findings obtained using established acoustic measures and findings from previous research.

1.6 Summary and Targets of the Current Study

For the present study we employed ultrasound tongue imaging to look at the perceptually fluent speech of people who stammer, which we compared against the perceptually fluent speech of people who do not stammer. The objective was to see whether even the fluent speech of people who stammer differs from that of typical speakers when looking at the articulatory level in order to get at the pervasiveness of disfluency at the global level – even when not perceptually salient.

We addressed this question by comparing articulatory patterns in the fluent speech of people who stammer to those of people who do not stammer. Because coarticulation was previously claimed to be problematic in the speech of people

who stammer (Belmont, 2015), transitions between consonants and vowels were of particular interest.

To understand disfluency, one needs to understand the fluent speech of the same people. Fluent and disfluent parts of speech are complementary components where one excludes the other. Understanding the fluent speech of a speaker therefore informs about the disfluent speech of that same speaker. For a full understanding of the fluent speech of PWS it needs to be investigated in comparison to the fluent speech of people who do not stammer. Whilst considerable work has been done to compare typical disfluent and pathological disfluent speech (Alfonso, 1991; Brocklehurst, 2011; Corley et al., 2007; Harbison et al., 1989; Hubbard & Yairi, 1988; R. Lickley, 2017; Postma & Kolk, 1993; A. I. Shapiro & Decicco, 1982; Shriberg, 1995; Yaruss, Newman, & Flora, 1999), less is known about the fluent speech of people who stammer (e.g., Max & Gracco, 2005; McClean & Levandowski, 1994; Wieneke, Eijken, Janssen, & Brutten, 2001; Zimmermann, 1980). In this thesis we seek to explore whether the perceptually fluent productions of people who do and do not stammer differ at the motor control / execution level.

Ultrasound tongue imaging was used to record images of the tongue surface during speech production of people who stammer. This technique is non-invasive while providing rich midsagittal information about the tongue moving in time and space. In contrast to other articulatory instrumentation ultrasound images can be recorded at a very high framerate providing insights into the kinematic nature of tongue movements. The high frame rate allows to observe even seemingly trivial tongue movement that might not always be apparent at the acoustic level. Ultrasound tongue imaging is particularly valuable for stammering research where speakers do not seem to differ from non-stammering speakers when speaking fluently.

Employing acoustic as well as articulatory data, both overt and covert speech events can be accessed to investigate the mechanisms underlying stammering. Building on Wingate's Fault-Line hypothesis (Wingate, 1988), we explored the transition from the syllable onset to the rhyme in the fluent speech of people who stammer.

Ultrasound images were recorded at the rate of 120 frames per second and used to extract kinematic information of the tongue movement in transition from syllable onset to syllable rhyme. From the kinematic information we extracted measures of articulatory duration, maximum velocity, and average speed as a ratio of duration and distance. We conducted an acoustic analysis using formant slopes and locus equations in order to determine whether differences apparent in the articulatory record could be identified from the acoustic analyses alone.

Throughout this thesis, we will try to answer the following questions:

a) General Research Questions:

- Does the fluent speech of PWS exhibit characteristics different to the speech of control speakers? If yes,
    - where can we localise these differences?
    - what measures express differently in the fluent speech of PWS?

b) Research Questions relating to the acoustic signal:

- To what extent does the perceptually fluent speech of PWS differ from that of PNS in CV transitions?
- Do acoustic measures taken from the fluent speech of PWS and PNS reflect the differences in timing found in previous literature?
- Do acoustic measures taken from the fluent speech of PWS and PNS reflect the differences in coarticulation found in previous literature?

c) Research Questions relating to the articulatory signal:

- Are articulatory measures coherent? / Do they appear to be valid?
- Do articulatory measures taken from the fluent speech of PWS and PNS reflect the results found in the acoustic data?
- Do articulatory measures taken from the fluent speech of PWS and PNS reflect the differences found in previous literature?

2 Method

The following chapter presents the frame within which the study was conducted. Information detailing participants (section 2.1), materials (section 2.2), instrumentation (section 2.3), and recording procedure including instructions (section 2.4) will be provided. Further, we will describe data screening (section 2.5) which was implemented to confirm that participants distinguished the three vowel conditions and to categorise perceptually fluent and disfluent productions to ensure that only the former were included in the subsequent analysis (chapters 3 and chapter 4). Recordings that were disregarded from the analysis due to perceptual disfluency are discussed qualitatively in section 5.3.

2.1 Participants

Twenty-one adults were recruited (13 adults who stammer and 11 control speakers) from the City of Edinburgh and surrounding areas using the research recruitment systems of Queen Margaret University and Edinburgh University as well as the British Stammering Association Network. Participants were sent information about the study. Information material differed slightly for the two participants groups (see Appendix F: Information Sheet for PWS and Appendix G: Information Sheet for PNS). Once participants volunteered to take part in the study, they were required to sign the consent form (see Appendix H: Consent Form).

All participants were 18 years of age or older at the time of the study. Participants did not report any hearing difficulty and had normal or corrected to normal vision. None of the participants reported neurological or cognitive impairments that could affect their participation in the study. Participants were compensated for their time and contribution with £15. The study was granted ethical approval by the Research Ethics Panel of the Queen Margaret University, Edinburgh.

Table 3 *Demographic information by participant group*

	PWS	PNS
age mean (SD)	34.1 (14.2)	33.4 (12.2)
gender: male / female	6 / 3	6 / 3
handedness:	9 / 0	8 / 1
education: A-Level / First / PG / NA	2 / 3 /3	2 / 2 / 5

People who stammer (PWS) were officially diagnosed (by a qualified speech and language therapist) as having a stammer and self-reported that the stammer commenced by the age of 10 years. None of the participants reported further speech impairments. While we recruited PWS, we sought to find suitable matched control speakers. Control speakers were matched for age, gender, and handedness as well as for educational background (highest academic degree achieved). Information to match speakers were obtained through an online questionnaire that was filled in prior to the recording (see the online questionnaire in section 7.9).

Two measures were employed to assess the severity of the stammer at the time of the study, which were further used to confirm the participants' categorisation as PWS. First, a standardised assessment tool for stammering, i.e., the Stuttering Severity Instrument (SSI) (Riley & Bakker, 2009) indicated the severity of the stammer based on quantifiable linguistic and extra-linguistic measures (see section 2.1.3). Additionally, a more subjective questionnaire was used to document the participants' perspective of their stammer based on the Overall Assessment of the Speakers Experience of the Stutter – a quality of life measure for PWS (OASES: Yaruss & Quesal, 2006 see section 2.1.3.2).

After exclusion of four PWS and two PNS (see section 2.1.2), data from nine PWS and nine control participants (PNS) were analysed (see Table 3 and Table 4 for demographic information on the speaker groups as well as the individual speakers).

2.1.1 English Language Ability

Though native speakers of English were preferred, we included bilingual speakers who acquired English before age six who were fluent in English preferably with British English pronunciation. All but one experimental participant were native speakers of English (including varieties of Scottish English, Irish English and one speaker of Pakistani English). Irish, Scottish and English accents of participants who stammer and control speakers were balanced. The non-native speaker of English (speaker 5) is native speaker of Polish who was balanced with a Polish control participant (speaker C) of same gender, equal age and educational background. The speakers' knowledge of English was enquired, and it was ensured that both speech perception and production were near-native.

2.1.2 Exclusion of Speakers

2.1.2.1 Exclusion due to Accent Variant

The language variant spoken by speaker 3 was Pakistani English. His production of stop consonants exhibited typical features of Pakistani English, with /p, t, k/ being realised in an unaspirated form when immediately preceding a stressed vowel. This pattern of realisation contrasts with that observed in speakers of British English, for whom /p, t, k/ are all realised in an aspirated form ([p^h], [t^h] and [k^h]) prior to a stressed vowel. As 'pə', 'tə' and 'kə' were amongst the key target syllables in the current study, the decision was taken to exclude speaker 3 from the analysis to minimise any risk of this accent variant impacting the potential to correctly identify group differences in realisation of the CV transition.

2.1.2.2 Exclusion due to Poor Data Quality

Ideally, the tongue surface structures would image as a white line clearly contrasting with other articulatory structures, such as the tongue body or the palate that would show in dark grey or black. Perfect image quality is admittedly rarely the case. For the splining of the tongue surface in motion (see section 4.1.1) it was necessary to see the tongue surface not only in lowered or medial position in the

vocal tract, but more importantly when raised and in palatal contact. The palate is the articulatory structure that is most distanced from the ultrasound probe which is why it is increasingly difficult to image the tongue surface the more raised / closer to the palate it is. The decision of whether data was found acceptable or not was based on the amount of tongue surface that could be imaged between the hyoid and the mandible and the ease with which the tongue contour could be tracked automatically (using AAA) when the tongue body was raised.

The quality of ultrasound data recordings varies according to individual speaker anatomy and physiology. Depending on the physical structures of the speaker, i.e., height of the oral cavity, bone structure and amount and quality of tissue underneath the mandible, the quality of the image varies. For the present thesis, the quality of the ultrasound image of a speaker determined its availability for articulatory analysis. In cases where the ultrasound recordings were not of sufficient quality the speakers' data needed to be excluded from further processing.

On this basis, data from 3 participants (PWS 4, 9 and 11) were judged to be insufficient to allow articulatory analysis. The small amount of tongue surface visible did not inform sufficiently about the tongue moving in the oral cavity. Data from these participants was therefore excluded from both articulatory and acoustic analysis, as was the data of participants from the control group who had been selected as matches for two of the excluded PWS (i.e., PNS F and G). The articulatory and acoustic analyses were therefore based on data obtained from nine PWS and their nine control speakers.

2.1.2.3 Final Participant and Control Selection

The nine PWS participants consisted of a group of six male and three female right-handed participants with a wide spread in age (age range: 20 to 60, mean age: 34.4, SD 14.2) comparable to that of the control speakers (age range: 20 to 60, mean age: 33.6, SD 12.2).

Table 4 *Demographic information on individual participants*

PWS	1	2	3	4	5	6	7	8	9	10	11	12	13
Gender	male	female	male	male	female	male	female	male	male	male	male	male	male
Age band	50-60	20-30	20-30	20-30	20-30	20-30	30-40	20-30	40-50	50-60	20-30	30-40	30-40
Handedness	right	right	left	right	right	right	right	right	right	right	right	right	right
Degree	PG	A-Level	A-Level	A-Level	First	First	First	A-Level	PG	no info	First	PG	PG
L1	English	Scottish	Urdu	Irish	Polish	Scottish	English	Scottish	Scottish	Scottish	Scottish	Irish	Scottish
PNS	A	D	G	F	C	H	I	E		B		K	J
Gender	male	female	male	male	female	male	female	male		male		male	male
Age band	50-60	20-30	20-30	20-30	20-30	20-30	30-40	20-30		50-60		30-40	20-30
Handedness	right	right	right	right	right	right	right	right		right		*left*	right
Degree	PG	A-Level	First	A-Level	*PG*	First	First	A-Level		PG		PG	PG
L1	English	Scottish	Scottish	Irish	Polish	Scottish	English	Scottish		Scottish		*English*	Scottish

2.1.3 Confirming the Stammer

2.1.3.1 Stuttering Severity Instrument

Both the reading task and the semi-structured interview were recorded in a soundproof studio at Queen Margaret University. The participant and the researcher were seated face-to-face at a distance of approximately 1.5 meters. A video camera was placed to the right of the researcher recording the speaker at an angle of approximately 160 degrees. The participant was seated in front of a monochrome wall to avoid any distraction in the video. The video captured the participants' upper body to ensuring that facial features were captured and available to the subsequent analysis.

Speakers were assessed on both the reading (see Appendix J) and the speaking task from the semi-structured interview. The severity of the stammer was scored and related to a severity equivalent to (1) the percentage of syllables stammered, (2) the average duration of the three longest disfluencies, and (3) physical concomitants. Total scores could range from about 10 points (very mild) up to 46 points (very severe). The video recorded material was assessed for frequency and duration using the Computerized Scoring of Stuttering Severity (Version 2; CSSS-2.0; see Table 5 for results). Physical concomitants were judged based on the recorded video as well as notes taken during the recording. All videos were recorded onto mini DVDs, which, it subsequently emerged, were highly prone to corruption. For that reason, the data from three speakers (PWS 2, PWS 3 and PWS 13) were lost and could not be retrieved. We were hence only be able to report test results for the remaining seven speakers who stammer. For speakers PWS 2 and PWS 13 (speaker PWS 3 was excluded from analysis) it was clear to the experimenter that a stammer was present in both speakers. Notes that were taken during the interview as well as audio recordings that were made independently from the video recording indicated that the two speakers presented with a mild-to-moderate stammer.

Table 5 *Stuttering Severity Instrument (SSI-IV) Ratings*

PWS	Reading	Speaking	Frequency	Duration	Physical	Sum	Equivalent
1	5	6	11	6	8	25	Moderate
2	-	-	-	-	-	-	Mild-to-moderate
3	-	-	-	-	-	-	-
4	5	6	11	6	8	25	Moderate
5	7	8	15	6	5	26	Moderate
6	4	7	11	4	4	19	Mild
7	4	5	9	2	2	13	Very mild
8	5	6	11	6	8	25	Moderate
9	5	8	13	4	8	25	Moderate
10	7	6	13	14	10	37	Very severe
11	2	5	7	2	3	12	Very mild
12	8	5	13	6	3	22	Mild
13	-	-	-	-	-	-	Mild-to-moderate

For the nine experimental participants in the study, we found that based on linguistic and extra-linguistic behaviours, the severity of their stammer reached from very mild in one speaker (PWS 7) over mild in two speakers (PWS 6, PWS 12), mild-to-moderate in two speakers (PWS 2, PWS 13) and moderate in three speakers (PWS 1, PWS 5, PWS 8) to one speaker whose stammer was categorised as very severe (PWS 10).

2.1.3.2 Overall Assessment of the Speaker's Experience of Stuttering

For the Overall Assessment of the Speaker's Experience of Stuttering (OASES), participants who stammer were required to fill in an online questionnaire (see Appendix I). The researcher was always available in case clarification was needed. In the OASES, questions are arranged in four sections referring to (1) rather general

information on the stammer, (2) the participants' feelings towards their stammer, (3) their feelings towards communication in daily life, as well as (4) to what extent participants felt the stammer affected their quality of life. For every question, participants were asked to respond on a 5-point Likert scale with the additional option of not providing a response. For each section, a relative severity was obtained by adding up points and relating them to the number of questions for which an answer was provided. An overall assessment was based on the average of the rating for the four sections (see Table 6).

Table 6 *Overall Assessment of the Speakers Experience of the Stutter (OASES)*

PWS	Section 1	Section 2	Section 3	Section 4	Total Score	Severity Equivalent
1	31.58	34.40	26.36	31.37	31	Mild-to-moderate
2	57.00	50.40	52.00	48.00	52	Moderate
3	61.00	81.60	75.79	68.70	72	Moderate-to-severe
4	57.00	50.40	52.00	48.00	52	Moderate
5	63.00	57.60	64.00	49.09	58	Moderate
6	51.00	42.50	42.96	30.00	41	Mild-to-moderate
7	73.00	76.80	74.40	63.20	72	Moderate-to-severe
8	63.75	44.00	38.26	28.80	42	Mild-to-moderate
9	70.00	54.40	21.67	28.80	43	Mild-to-moderate
10	55.79	59.20	76.84	55.79	62	Moderate-to-severe
11	34.00	35.20	30.83	20.00	30	Mild-to-moderate
12	64.00	53.60	49.17	38.40	51	Moderate
13	38.00	23.20	33.91	25.60	30	Mild-to-moderate

For the nine experimental participants in the study we found that the experience of their stammer reached from mild-to-moderate (in four speakers: PWS1, PWS6, PWS8, PWS13) over moderate (in three speakers: PWS2, PWS5, PWS12) to moderate-to-severe (in two speakers: PWS 7, PWS10).

2.1.3.3 Summary of Severity Assessment

Looking at the two methods of assessing the severity of the stammer, one could observe differences in how the two ratings categorised the severity of the stammer. While participants' stammers were classified at a wider spread, covering categories from very mild to very severe in the SSI, the same speaker's stammers were classified in only three categories spanning from mild-to-moderate to moderate-to-severe using the OASES.

For most speakers, the ratings from the two assessment tools returned the same category (PWS 4, PWS 5) or the classification differs only by half a category (PWS 1, PWS 6, PWS 8, PWS 9, PWS 10) which might be owing to the difference in available categories. For two speakers, however, the categorisation of the stammer differed meaningfully (PWS 7 and PWS 12) depending on the assessment tool.

The reason for that difference in severity categorisation (see figures in Table 5 and Table 6) may be owing to differences in what is assessed or owing to differences in how the stammer is assessed. While the SSI assessment tool categorises the stammer, the OASES tool assesses the effect of the stammer based on the participant's experience of the stammer. Further, the SSI bases the severity categorisation on linguistic and extra-linguistic measures such as stammering frequency, duration of stammer event and physical concomitants – all perceptually available to the interlocutor. What is not accounted for, however, is the amount of covert instances of stammering that are only available to the speaker himself.

For speaker PWS 7, for example, the SSI and the OASES returned almost the opposite severity categories with "very mild" on one hand and "moderate-to-severe" on the other. The example of that speaker showed how much the

perception of others and the perception of self can differ with regards to stammering. A possible explanation as to why speaker 7 experienced his stammer as more severe may stem from (amongst others) the fact that perception of one's own speech includes covert as well as overt symptoms of stammering whereas traditional assessments include overt symptoms only.

Despite differences in what was measured and how it was measured both instruments confirmed that speakers presented with a stammer at the time this study was conducted.

2.2 Materials

The data presented in this thesis were collected in the context of a larger project. The focus of the thesis was the articulatory realisation of single words. How words are realised in running speech is subject to many variables, including psycholinguistic variables such as word frequency and concurrent processing requirements. For this reason we employed consonant-vowel (CV) phrases and focus specifically on theoretical understandings pertinent to this context.

Materials were designed in a highly experimental manner to allow for a maximally focussed investigation of the transition between consonant (C) and vowel (V). During data collection, participants produced consonant-vowel (CV) utterances in isolation and in carrier phrases. Each target CV was initiated with a /ə/ (prothetic vowel (VP); as introduced by Austin, 1941; Willis, 2006; see section 2.2.3). The CV syllables consisted of onsets /p, t, k, s, ɔ/ and rhymes /ɑ, i, ə /. Onsets and rhymes differed to investigate the role of varying place and manner of articulation. Utterances were produced in typically voiced speech and in whisper. All participants produced all V_PCV utterances in all conditions. Data were collected in two recording sessions on the same day where sessions were balanced for mode of speech (i.e., utterances produced in typically voiced speech and in whisper).

2.2.1 Stimulus Modi

Table 7 *Modes of stimulus production*

Session	Condition	Voicing [+/-]	Carrier Phrase [+/-]
Session 1	typically voiced	[+ voice]	[- carrier phrase]
	in whisper	[- voice]	[- carrier phrase]
	in carrier phrase	[+ voice]	[+ carrier phrase]
Session 2	typically voiced	[+ voice]	[- carrier phrase]
	in whisper	[- voice]	[- carrier phrase]

The material was randomised for each speaker and condition (typically voiced and in whisper). Using a 'Latin Square' paradigm, we ensured that the two speaker groups (PNS and PWS, see Table 7) produced the same materials. We further controlled that different speakers produced the same materials typically voiced and in whisper.

Half of the typically voiced and whispered material was produced in the first session, followed by the other half in the second session. Utterances produced in carrier phrases were produced as part of the first recording session. For both recording sessions, acoustic as well as kinematic labial and kinematic lingual data were obtained. The time between ultrasound recorded sessions was used for the OASES and SSI assessments (Riley & Bakker, 2009; Yaruss & Quesal, 2006) lasting approximately 30 minutes. In this thesis we examined exclusively acoustic and lingual recordings of V_PCV utterances featuring the onsets /p, t, k, s/ produced in isolation in the typically voiced condition – unless otherwise specified.

2.2.2 CV Syllables

Table 8 *Consonant-vowel composition*

manner of articulation	place of articulation	stimulus	[+ target vowel]	[- target vowel]
C = plosive	bilabial	/p/	/ɑ/, /i/	/ə/
	alveolar	/t/		
	velar	/k/		
C = fricative	alveolar	/s/	/ɑ/, /i/	/ə/

Stimuli consisted of V_PCV utterances consisting of a prothetic /ə/ (V_P) followed by the target CV syllable with different combinations of syllable onsets (C) and syllable nuclei (V). Syllable onsets were manipulated to include different manners of articulation (plosive = /p, t, k/ and fricative = /s/) as well as places of articulation (bilabial = /p/, alveolar = /t/ and /s/, velar = /k/). In addition to the manipulation of syllable onsets, syllable nuclei differed in height including high (i.e., /i/) and low vowels (i.e., /ɑ/) with relatively peripheral tongue displacement as compared to a rather neutral articulatory target (i.e., /ə/) as can be seen in Table 8. Speakers were asked to use a schwa-like central vowel that would contrast with the cardinal vowels where the tongue is advanced or retracted. Hence, articulatory effort would be expected to be greater for corner vowels when compared to the neutral articulatory setting for schwa.

The resulting material consisted of high-frequency CV utterances. While these CV utterances consisted of non-words, their high-frequency nature implies stable feedforward models similar to those of real words (Kröger, 2013; Max et al., 2004).

2.2.3 Prothetic Schwa

Speakers were instructed to produce a prothetic schwa (V_P) preceding every CV syllable. In a short training session immediately preceding the experiment we provided participants with examples of monosyllabic nouns with indefinite articles (i.e., 'a shoe', 'a door') and instructed participants to employ the same pattern when producing the stimuli consisting of V_P + CV target syllable where V_P was produced similar to the indefinite article in the training prompts. Participants were asked to produce up to 10 test stimuli. Once the participant felt comfortable producing the prothetic schwa and the following CV syllable, the training session was concluded, and the actual recording was prepared.

The schwa sound serves as a prothesis to fulfil several purposes:

- Maintaining stable stress pattern
- Steady and consistent starting position of the tongue,
- Preventing participants from bracing the tongue against the roof of the month, and
- Inducing fluency.

2.2.3.1 Stress Pattern

During data recording we controlled for stress, as it has been claimed to play an important role in disfluent speech (Wingate, 1984b). Linguistic stress refers to the relative emphasis that is put on a syllable. Greater stress is phonetically realised by an increase in either or both amplitude and vowel length. Stress is further associated with the full (as opposed to reduced) articulation of the vowel (Crystal & House, 1990; Lehiste & Peterson, 1959; Lindblom, 1963).

Wingate and others (Au-Yeung et al., 2003; Dworzynski et al., 2004; Hubbard & Prins, 1994; Natke, Gosser, Sandriseser, & Kalveram, 2002; Prins, Hubbard, & Krause, 1991; Wingate, 1984b) have claimed that there is a link between stress and the increased occurrence of disfluencies. According to them, stressed syllables are

generally more likely to be disfluent than unstressed syllables. Stress could therefore function as a predictor for disfluencies. The studies from Wingate and others did, however, draw their findings from mainly acoustic studies, which means that claims were mostly limited to overt (perceptually salient) dysfluencies. Taking it one step further, it may be plausible that linguistic stress may function to not only predict an increase in the number of 'overt' but also of 'covert' occurrences of stammering (including uncategorised lingual behaviours deviating from those of people who do not stammer). With this potential effect on the fluency of speech in mind, we controlled for stress and instructed participants to produce all target syllables with lexical stress.

To increase the probability that participants put the stress on the target syllables we instructed participants to produce every target syllable with a preceding prothesis for which we chose schwa. We chose schwa because it is typically unstressed and because it is not typical for English to have two stressed or two unstressed syllables in juxtaposition. Preceding the target syllable with an unstressed schwa (V_P) and introducing the participant to the iamb stress meter in the training phase, ensured that speakers were consistent at producing the stress on the target syllable. The iamb meter was working in two directions: In addition to imposing the indented stress on the target syllable, the iamb meter also affected the schwa in that it reinforced its reduction thereby not letting it slip towards an open [ʌ] or [ɪ] which SSE speakers tend to use in unstressed position (Abercrombie, 1979, 1991; Durand, 2004, p. 94).

2.2.3.2 Steady Starting Position

Employing a prothesis that was produced prior to the CV target syllable further led to a stable and consistent articulatory starting position for the CV syllable. The movement into the consonant was therefore comparable across target syllable and speaker. Without a consistent starting position, movements into the consonant would differ depending on the starting position and this would further add to the already high complexity of dynamic analysis of articulatory movements.

Schwa was chosen to precede the target syllable because it is a central vowel that is "produced with a neutral setting of the articulators [requiring] no displacement of the articulators from the neutral position" (Giegerich, 1992, p. 68). This is useful as it implies that the lingual configuration is independent of the size and shape of the speakers' vocal tract. The neutral starting position further implies that there is minimal articulatory influence (i.e., co-articulatory effect) on the actual target syllable (Gick, 2002; Watkins, Baptista, & Watkins, 2006). The neutral nature of the pre-stimulus token schwa reduces co-articulatory effects onto the target syllable to a minimum, which then allowed us to observe and compare maximally pure movements into the target syllable across stimuli and speakers.

2.2.3.3 Preventing Bracing Behaviours

A side effect of the neutral starting position on schwa is that it prevents speakers from starting off from a bracing position (Gick, Allen, Roewer-Després, & Stavness, 2017; Gick, Allen, Stavness, & Wilson, 2013; Stone & Lundberg, 1994). Bracing describes a speaker's behaviour to press his or her tongue against the palate – a behaviour often observed during speech preparation in experimental setups. Speakers vary in their preference for an articulatory rest position. While some speakers prefer a lowered tongue, others prefer to brace, i.e., by pressing their tongue against the palate.

With ultrasound imaging technology bracing is problematic as it is nearly impossible to capture the tongue when resting at the palate. Pressing the tongue against the palate reduces any air pockets above the tongue surface that are necessary to image the tongue surface. Bracing behaviour would therefore render movements into the CV target stimulus indiscernible making it difficult to impossible to analyse. This would in effect lead to the exclusion of substantial amounts of data. Inserting a schwa before the actual target syllable was therefore helpful as it generated data that was accessible using ultrasound which could then be included in the analysis.

2.2.3.4 Inducing Fluency

Based on the experimenter's perception, speakers who stammer were found to be more fluent in the syllable production task when compared to the reading or conversation data from the stammering severity instrument (Riley & Bakker, 2009) where they appeared to be more disfluent. Despite noticeable individual variation, PWS showed an overall increased fluency in the less natural, i.e., experimental, condition which may be owing to how we designed the prompts we used for the present study.

Previous studies found that certain conditions can affect fluency in the speech of PWS (Andrews, Howie, Dozsa, & Guitar, 1982). Evidence for increased fluency could be found when inserting vowels preceding the onset of speech by people who stammer (Dayalu, Saltuklaroglu, Kalinowski, Stuart, & Rastatter, 2001). This effect could be comparable to lip aperture data that showed pre-speech movement in places where the movement would not be expected which we will be briefly discuss in the subsequent section.

2.2.3.5 Preparatory Mechanism

The following section is to pre-empt doubts about the CV nature of the prompts we asked participants to produce. Participants were asked to produce a prothetic schwa prior to each CV prompt resulting in V_P#CV utterances. The prothetic schwa induces a release before moving towards the C target. For the pre-speech release the speaker opens the lips and places the tongue central to the oral cavity. From there the articulators move towards the target setting where, for example, lips are closed for bilabial stops and the tongue is placed at the palate for palatal stops. The pre-speech release therefore allows investigation of movements into the closure for the consonant.

Pre-speech release appears to function as a preparatory mechanism for speech production that is natural to speakers. It was previously observed by Scobbie (personal communication 27-04-2018) who explored labial pre-speech movements

(lip opening prior to speech initiation) in a study where participants were asked to produce words commencing with a consonant where C = bilabial /m/ or /p/ immediately followed by a vowel in two conditions – a neutral condition and a clear speech condition.

Scobbie observed labial release following prompt display but prior to production of the target word. While lips were typically closed at the time the prompt was displayed, speakers tended to release (lip opening) prior to bilabial closure in word initial position. While this pre-speech release is not physically required to produce a bilabial stop, it appears to be a natural preparatory mechanism for the initiation of speech, observed in all speakers (to different degrees).

2.2.4 Stimulus Repetitions

Table 9 *Stimulus repetitions*

/ə/ + /p, t, s, k/ + /ɑ, i, ə /	12 CV combinations
12 repetitions	144 utterances

Combining the four syllable onsets (/p/, /t/, /k/, /s/) with three syllable nuclei (/ɑ/, /i/, /ə/) resulted in 12 different CV target stimulus combinations each preceded by a /ə/). The 12 sets of combination were produced 12 times in randomised order. This yielded a total of 144 utterances produced (see Table 9). Following the recording, disfluent productions were excluded from analysis. The number of repetitions was sought to ensure sufficient data to test statistical significance while avoiding effect of fatigue in participants, which is a general concern in not only clinical studies (Leisman, Zenhausern, Ferentz, Tefera, & Zemcov, 1995; Sawyer, Chon, & Ambrose, 2008; A. Smith, 2006).

2.3 Instrumentation

2.3.1 Ultrasound Tongue Imaging

The recordings were made in the speech laboratory of the Clinical Audiology Speech Language (CASL) Research Centre at Queen Margaret University. The laboratory that was used for the recordings consists of two smaller sound-treated studios. Participants were seated in front of a computer screen in one of the studios. The ultrasound probe and a small microphone were attached to a headset (Articulate Instruments Ltd, 2008) that participants wore during the recording session. The headset was used to stabilise the ultrasound probe and the attached microphone ensured clarity of sound due to the proximity to the speakers' mouth.

The ultrasound machine was remotely controlled via Ethernet from a PC in the neighbouring studio, running Articulate Assistant Advanced software (Wrench, 2015) version 2.14 and 2.15. The researcher in the neighbouring room controlled the beginning and the end of each recording manually. As soon as the recording was initiated by the researcher, a fixation cross appeared on a green background for 300ms. Following the 300ms delay participants perceived a beep sound cueing them to read the prompt that appeared simultaneously on the screen.

Audio and ultrasound signal were recorded of participants reading the stimuli off the computer screen. Both signals were then sent to the controlling PC in the neighbouring room where they were synchronised. The ultrasound machine that was used to record the data was an Ultrasonix SonixRP machine, which has the advantage that it is particularly precise when synchronising data.

The ultrasound probe was a micro-convex type that recorded at 121fps with 63 scan lines evenly spread over a 135-degree field of view (FoV). The depth was set to 80 mm and the echo return vectors had 412 samples resulting in a resolution of approximately 5 pixels per mm. The transducer frequency was 5MHz providing an axial resolution of approximately 0.9 mm. Every time an ultrasound scan consisting

of the 63 scan lines had been recorded, a pulse was generated. This pulse was then used as the synchronising signal and sent to the same sound card as the audio signal. Articulate Assistant Advanced software (Articulate Instruments Ltd, 2012, 2014) used this synchronisation signal to assign each complete scan image an exact time point.

2.3.2 Data Orientation

The recording of the ultrasound sessions lasted approximately 30 minutes, each ranging from 25 to 35 minutes. The actual recording duration depended on the speed at which the participant produced the stimuli, which also directly affected the duration it would take the system to store the data following the recording of each prompt. Over the duration of the recording we aimed for a maximally stable ultrasound image, which could not be achieved by a hand-held probe alone because even slightly shifting the probe location on the participants' chin or slightly changing its angle would affect the image considerably. Keeping it in a stable position was therefore crucial for later analysis.

2.3.2.1 Head Set

To achieve a steady image of the ultrasound tongue imaging all participants were fitted a stabilisation helmet (Articulate Instruments Ltd, 2008) to which the ultrasound probe could be attached. The headset is used to hold the ultrasound probe in place. This way, speakers could move their heads naturally with the ultrasound probe remaining in a stable position relative to the participant's head. The headset ensures that the information recorded is consistent and stable despite the speaker's head movement. It reduces noise that would otherwise falsify any comparative, especially quantitative analysis (Zharkova et al., 2015). It is a safe and relatively non-invasive procedure with negligible impact on the speaker's articulation (Villegas, Wilson, Iguro, & Erickson, 2015).

The headset is made up of an aluminium construction weighing 0.8 kg. To distribute the weight across the scalp the headset is lined with gel and neoprene padding. The

headset can be adjusted to participant's head size on various dimensions. Once the headset is fitted to the participant's head the ultrasound probe is inserted into a clamp underneath the participants' chin and fastened on the mid-sagittal plane. To ensure a good image, the fastened probe can rotate or translate on the mid-sagittal plane. Translational movement of the probe may be necessary to position the transducer of the probe at the soft tissue underneath the tongue body between the hyoid bone and the mandible. Pitch rotation is required to rotate the probe, so as to capture the image of the tongue central to the shadow of the hyoid and the mandible. A central orientation to these two bone structures ensures that maximal information of the tongue surface is captured. The headset would control for probe movement within a session.

To control for movement of the ultrasound probe within and between recording sessions and to allow for within- and between-speaker comparison ultrasound images need to be aligned. Typically, ultrasound images consist of black and white pixels that represent the reflection of the ultrasound beam from organic structures of the individual. These organic structures differ in shape and size for each individual, which means that they are not suitable as referent for aligning data. In addition to the represented structures, we employed two measures that would add referents to the ultrasound signal, which then served for alignment of recordings within and across speakers.

2.3.2.2 Palate Trace

First, we consider the palate trace. Anatomical features of humans differ greatly between individuals (Fant, 1966; Fuchs et al., 2006; Rudy & Yunusova, 2013), even adults. Research has found that children are more likely to have a smaller oral cavity and that the size and the shape of the oral cavity changes with increasing age. What has not yet been established, however, is the relation between body size and the size of the oral cavity. Though there may be that tendency, individual differences require us to have a closer look at the palates of participants to understand potential effects on articulation.

The hard palate does not move when people speak. As such it can be used as a reference point indicating the maximum movement possible of the tongue. More importantly, it can be used as reference when checking whether the probe has remained stable over a recording session, which is a prerequisite to compare and overlay tongue images that were collected throughout a recording session.

For the present study, we use the palate as a reference point to overlay images collected within a recording session as well as across the two sessions recorded by each speaker. As mentioned earlier, recording sessions lasted up to 35 minutes, which gave time for the headset (and with it the probe) to move. Between sessions participants took off the stabilisation headset, which then needed to be reattached for the second ultrasound session. Only slight shifts in the probe orientation would result in data that could not be overlaid without correction using a common reference point. The trace of the hard palate is therefore useful in two ways: it allows controlling for movement of the probe within a session as well as across several sessions.

The question remaining is how we get the trace of the palate. It sounds fairly simple and straightforward: Because the ultrasound image captures the tongue surface (i.e., tissue meeting air) we ask participants to swallow a sip of water whereby the tongue is sliding along the palate. The ultrasound image then captures the small amount of air that is trapped between the tongue and the palate resulting in a good palate-shaped trace of the tongue surface. A problem with this is, however, that the signal of the ultrasound probe may not be strong enough to image the tongue when it is at the maximum distance from the probe. For our recordings, the maximum distance for the signal was set at 8 cm, which was the trade-off with image quality.

2.3.2.3 Bite Plate

Another measure to control for ultrasound probe movements is the bite plate (Lawson, Stuart-Smith, Scobbie, & Nakai, 2015). In order to establish the bite plane, a plastic plate is used that measures approximately 6 by 10 cm. The plastic plate is inserted into the speaker's mouth where it is fixated by biting on the plate with the

molars. The plastic plate on its own does not show on the ultrasound image participants. To obtain the bite plane, the speaker presses his or her tongue against the underneath of the plastic plate, which generates a straight line in the ultrasound image (Scobbie, Lawson, Cowen, Cleland, & Wrench, 2011).

About 4.3 cm from the edge of the plastic plate is a fixation point. This fixation point is meant to press against the participant's incisors, indicating that 4.3 cm of the plate reach into the oral cavity. The resulting image shows 4.3 cm of flat tongue surface, which is where the tongue presses against the underneath of the bite plate. At the end of the 4.3 cm the tongue bulges and takes on its more natural curvy shape. The bulging of the tongue surface is useful as it indicates the end of the bite plate reaching into the oral cavity as far as 4.3 cm from the fixation point at the incisors.

The horizontal line / bite plane uses the speaker's teeth as a reference, which means it is independent from the speaker's posture, height, or oral cavity size. The ultrasound image on its own displays the tongue surface and muscular structures without internal referents for orientation to which the recordings of multiple speakers could be aligned. Establishing a horizontal plane in the ultrasound image allows aligning and comparing the images within and across participants.

2.3.2.4 Summary

Both measures, the palate trace and the bite plate in combination, provide information to enable us to compare data within and across sessions of a single speaker, and with some limitations to also compare across speakers. The palate trace is useful to capture the features of the individuals' oral cavity, such as its height, its shape or overall size of the alveolar, post-alveolar and in some cases even velar area. The bite plate, in addition, attempts to establish a common referent across participants. Relying on the bite plate alone to compare across speakers, however, needs to be done with caution as the bite plane uses the individual's teeth as referent to produce the horizontal line. While teeth are helpful to establishing

the horizontal plane as a referent to account for head rotation, it may still be subject to individual anatomic differences.

Establishing individual features is useful to compare images of the same participant within or across sessions; establishing something that all ultrasound images have in common is indispensable when attempting to investigate static or even dynamic tongue contours across participants. For the current study, we will draw on both of these data extrinsic measures to ensure stability of data rotation within and across session of individual participants. For the comparison across speaker, we will introduce a data intrinsic linear measures to characterise dynamic gestures through tracking the distance of the tongue surface from the probe along one of the measurement radii (see section 4.1 below).

2.4 Recording Procedure

2.4.1 Instructions on the Experiment

Participants were aware that the study investigated articulatory differences between people who stammered and people who do not stammer. Looking at developmental speech impairment we had to consider that participants would have acquired techniques to control their stammer. While some techniques might be controlled consciously, others might be automatized to the extent that they cannot be controlled consciously.

For the same reason we decided to not instruct participants on whether or not to use acquired techniques to speak as fluently as possible. We also did not instruct the participant to produce the utterances at a certain speed. Aiming at most natural data (whatever the level of stammer) we avoided raising the topic and drawing the participants' attention to the issue.

In an initial conversation with the participant, the procedure of the stammering assessment and the recording of the data were explained. Participants were

informed about the procedure of the study. They were informed that, over two 30-minute sessions their ultrasound, audio and video data would be recorded.

Participants were instructed on both the linguistic part of the study as well as the more technical part of it. For the linguistic information, the concepts of schwa as well as that of whispering were explained. Subjects were informed about the target tokens comprising V_PCV syllables where C covers the sounds /p/, /t/, /k/, /ɔ/ and /s/ that was combined with three different V comprising of the sounds /ɑ/, /i/ and /ə/. For the first vowel V_P that preceded the CV target syllable, participants were instructed to pronounce it like an indefinite article preceding a monosyllabic noun. Monosyllabic nouns with an indefinite article (e.g., “a shoe”, “a three”, “a four”) were provided to help participants understand the structure of the target utterance. Following the description, participants were asked to apply the same pattern to ten to twenty test stimuli. The training was discontinued at the point where both the experimenter and the participant felt confident that the materials and the pronunciation patterns were understood.

In addition to the linguistic instructions, a brief introduction to ultrasound recording was provided. This involved the general information about ultrasound being non-invasive and requiring water-based ultrasound gel as a conductive medium through which the ultrasound waves travel. Moreover, participants were introduced to three ways of controlling for head movement (see section 2.3.2). We ensured that participants fully understood the procedure as well as any explanations and instructions provided prior and during the recording of data. Participants were encouraged to ask questions at any time. A consent form was filled in and signed prior to participation in the study (see Appendix H).

2.4.2 Experimental Setup

For the presentation of the stimuli participants were seated in an adjustable chair in front of a computer screen. Stimuli were presented one by one on a computer monitor. Ultrasound tongue imaging data, acoustic data as well as video data were

recorded using AAA software. Ultrasound tongue imaging, audio and video recording are initiated together with the display of a fixation cross on the computer screen. The fixation cross remained stable for the duration of approximately 300ms, at the end of which the fixation cross disappears and the stimulus appears. Together with the display of the stimulus participants perceive an acoustic signal cueing them to articulate the stimulus. Participants were instructed to read the stimuli at their own pace.

The duration of the recording was managed manually, because automatically pre-set recording duration could lead to a loss of data when, for example, speakers produced the stimuli with a longer response time. The researcher was seated in a neighbouring room supervising the participants' utterance production from the acoustic signal as well as ultrasound and video image. The aim was to ensure that the material was fully recorded on all three channels, i.e., acoustic, labial and lingual data. Research has established that lingual articulatory duration outlasts the acoustic speech signal as well as labial articulatory movements (Bell-Berti & Harris, 1981; Krakow, 1999). Complete lingual movement can therefore be assumed to indicate completed acoustic speech signal as well as completion of labial motion for the transition of the respective CV syllable. The ultrasound image shows the lingual closure and release phase for each consonant. Recording duration is based on the ultrasound image showing lingual closure and release phase for each consonant. The researcher discontinued the recording as soon as the tongue reached a stable position following the release phase.

Once the speaker was prompted to produce the target stimulus, acoustic and ultrasound data were recorded and synchronised automatically during the recording. The video of the speaker's lip movement needed to be synchronised in a subsequent step. During the recording, a BrightUp device (Articulate Instruments Ltd, 2010) superimposes a white mark on the video, which can later on be used for synchronisation. Whenever the synchronisation device fails to superimpose the white mark, the researcher is notified as the recording cannot be synchronised and

is therefore not available for further video analysis. Every time the synchronisation failed the prompt was recorded again. Stimuli were also repeated if mispronounced or disfluent. The additional recording was meant to ensure a minimum of eight fluent productions of each prompt and speaker that would be available for analysis. To avoid any training effects the repeated recording did not follow immediately but with a delay of at least three prompts.

Participants were recorded over two sessions, using the stabilisation helmet for the ultrasound probe. The stabilisation helmet for the ultrasound probe can become uncomfortable over time, which led to a maximum recording time of approximately 30 minutes for each recording session. Both sessions of a speaker were recorded in one day.

2.5 Data Screening

2.5.1 The Three Vowel Conditions

As mentioned before, stimuli consisted of CV syllables with a combination of three different vowels. Vowels included the high vowel /i/, the low vowel /ɑ/ and the mid vowel /ə/. Having included Scottish speakers in the study, we were wary that Scottish speakers do not usually have the schwa-like target in their vowel inventory (e.g. in NURSE lexical set, in CV open syllables, unlike non-rhotic English accents in words like fir, her, purr).

To make sure speakers distinguished the three different vowels /ɑ/, /i/, and /ə/ in the nucleus of the stressed CV syllables, we carried out two types of analysis. First, we plotted the first and second formants taken at the mid vowel for all speakers combined. From the resulting plot (see *Figure 11*), we could see a clear divide between the high vowel /i/ on one hand and the mid and low vowels /ɑ/ and /ə/ on the other. The latter two, however, showed a large amount of overlap in their F1 and F2 distribution.

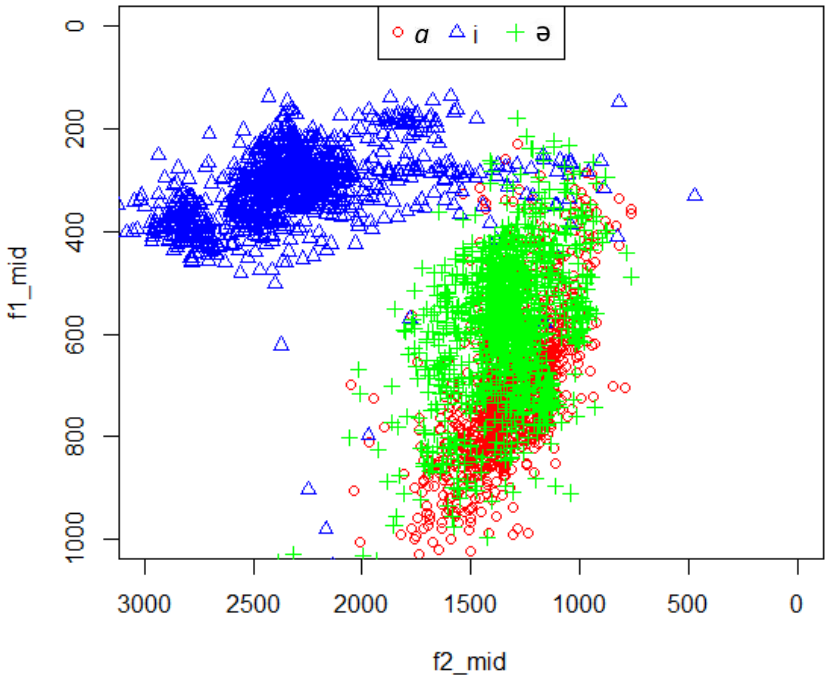


Figure 11 Vowel chart displaying /i/ (blue), /ə/ (green) and /ɑ/ (red) in /k/ context

To further see how similarly the mid vowel /ə/ and the low vowel /ɑ/ were produced and whether there was indeed no differentiation made, we had a closer look at the distribution patterns of F1 and F2 for the three vowels using a density plot which uses the density of observations to assign each region to a probability level. In contrast to the initial vowel chart (see *Figure 11*) that was taken across all participants, we further distinguished between male and female participants to obtain a more detailed picture including information on the distribution of observations (*Figure 12*).

The density plot (see *Figure 12*), like the vowel chart above (*Figure 11*), shows a clear distinction between the high vowel /i/ and the low and mid vowels /ɑ/ and /ə/. Further, the density information provided in the density plot, shows a clear distinction between the mid vowel /ə/ and the low vowel /ɑ/ which appears to be a reflection of gender specific F1 and F2 ranges (see *Figure 12*). Both groups of male

(left panel) and female speakers (right panel) show different preferences for where they produce vowels. The vowel /i/ is clearly distinct with a lower F1 and a higher F2 compared to the other two. The vowels /ɑ/ and /ə/ in comparison can be seen to partially overlap in their distribution; especially on F2. They can, however, be distinguished via the distribution along F1 where /ə/ is produced at overall lower F1 values when compared to /ɑ/ also reflected in the lower distributional F1 centres.

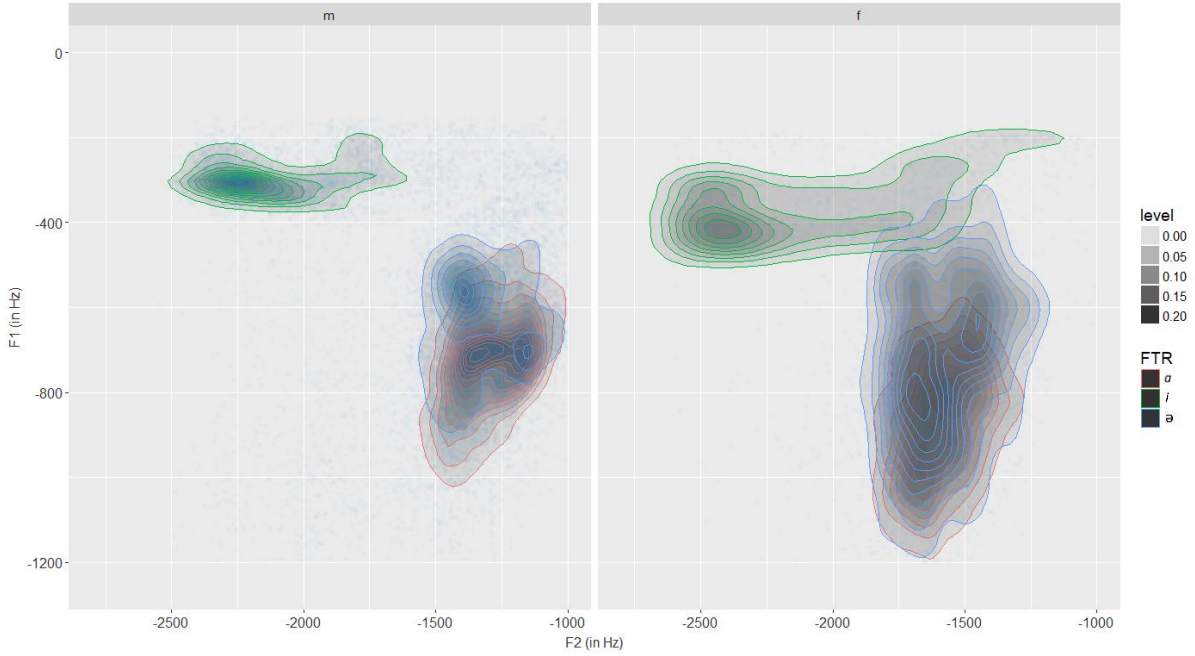


Figure 12 *Density plot displaying the density of data points for /i/ (green), /ə/ (blue) and /ɑ/ (red) in /k/ context by gender (male: left panel; female: right panel)*

These distributional centres appear to be rather straight forward for the male speakers. For the female speakers, however, we observe a somewhat bimodal distribution for the high vowel /i/ as well as the mid vowel /ə/ which might reflect differences among speakers, since the context remains the same. Overall, we find the mid vowel /ə/ to inhabit a lower F1 (approx. 550Hz for female; 680Hz for male) compared to the low vowel /ɑ/ (approx. 730Hz for female; 850 for male).

In addition to these rather qualitative impressions about the divide of /ɑ/ and /ə/ we have taken another more quantitative / automated measure to investigate whether the participants have produced /ə/ when instructed to do so. For this automated approach we used the FAVE (Forced Alignment and Vowel Extraction) program suite (Rosenfelder, Fruehwald, Evanini, & Yuan, 2011), an online interface for automatic speech alignment and vowel extraction. This tool is designed to insert phoneme boundaries based on F1 and F2 information from the acoustic signal of, amongst others, English speech data. Once segmented, it matches strings of phonemes against an in-built dictionary. The in-built dictionary uses the phonetic transcription code ARPAbet which includes information about consonants, vowels and stress. Stress is indicated using digits following the vowel where '0' represents no stress and '1' and '2'represents primary and secondary stress. While FAVE offers a dictionaries with words, it can be extended manually to also recognise non-words.

Table 10 *Entries added to FAVE aligner*

	/ɑ/	/i/	/ə/
/ɹ/	AA0	IY0	AH0
	AA1	IY1	AH1
/p/	PAA0	PIY0	PAH0
	PAA1	PIY1	PAH1
/t/	TAA0	TIY0	TAH0
	TAA1	TIY1	TAH1
/s/	SAA0	SIY0	SAH0
	SAA1	SIY1	SAH1
/k/	KAA0	KIY0	KAH0
	KAA1	KIY1	KAH1

We added non-word CV entries to the in-built dictionary to run the aligner on the material we collected for the current study. For each of the prompts included in the

study material, we added a separate entry. This resulted in 30 entries consisting of a combination of the consonants (including glottal stop) and the three vowels carrying primary stress as well as no stress (see Table 10).

We used the FAVE aligner to explore how well the instructed prompt and the produced prompt matched. The instructed prompt would be the CV utterances with varying consonant (x5), vowel (x3) and stress (x2). Looking at the data and having it aligned automatically, we were interested to see whether the same distinctions were present. As was expected, for prompts where V = /i/ the automatic aligner found a large correlation between the 'instructed' prompts and the actual realisations of that prompt transcribing them with 'IY' in 99%. For the mid vowel /ə/ and the low vowel /ɑ/ the discrimination was expected to be less distinct when compared to the discrimination of the high vowel /i/. To our surprise, however, FAVE correctly identified over 95% of the /ə/ and /ɑ/ sounds transcribing them as 'AH' and 'AA' respectively.

The clear automatic discrimination stood in stark contrast to the overlaps in distribution shown in the vowel chart (*Figure 11*) and the density plot (*Figure 12*) which provides a more detailed picture of the distribution and overlaps in distribution. Though the vowel charts gave the initial impression that speakers produced only a high vowel and non-high vowel contrast, the clear discrimination by FAVE indicates that individual speakers did distinguish between all three vowels. While the effect may not be perceptually salient, its acoustic availability justifies the distinction in the subsequent analysis.

An explanation to this apparent contradiction may be the fact that we would run the aligner FAVE on the data of one speaker at a time. The FAVE aligner uses the data that it is supposed to align as training data at the same time. This means that the batches of data that are fed in together also affect the algorithm with which that same batch is being aligned. Feeding in the data of each speaker separately also means that the aligner applied a slightly different algorithm to each speakers' data, which may account for the high discrimination rate for the mid and the low

vowels. What the FAVE aligner shows us is that the fact that the two vowels cannot be discerned in the pooled data from all participants (see *Figure 11* and *Figure 12*) does not mean that the individual speaker does not make the distinction.

Both the qualitative vowel plot and the density plot as well as the automated investigation using FAVE together suggest that speakers treated all three sounds as separate entities also allowing us to treat them separately in the following analysis.

2.5.2 Fluency Judgement

The objective of the current study was to investigate the fluent speech of people who stammer to explore whether even apparently fluent speech of PWS differs from the fluent speech of PNS. Investigating fluent speech in PWS, we needed to ensure that only perceptually fluent tokens (which could also include productions of covert stammering) would be included in the subsequent analysis. To perform the analysis on the fluent data, we categorised the recordings of PWS into fluent and disfluent speech. We undertook two approaches to remove overt disfluencies from acoustic and articulatory data.

Typically, the speech of people who stammer is defined via its disfluent characteristics. The disfluent recordings will be included in the discussion of this thesis (see section 5.3) where examples of disfluent recordings will be described qualitatively to highlight the highly variable nature of the manifestations of stammered speech.

2.5.2.1 Fluency Judgement of Acoustic Data

The acoustic data was categorised into fluent and disfluent recordings drawing on the perceptually available information. The experimenter (thesis author) listened to all recordings of PWS to distinguish acoustically fluent from disfluent recordings. Recordings for each PWS speaker were concatenated using the concatenate function in the speech analysis software Praat (Boersma & Weenink, 2015). The concatenated files contained recordings of the same prompt produced by an individual speaker.

This way recordings of a prompt were compared against other recordings of that same prompt. Judgements were made for a speaker at a time to decrease the influence of individual characteristics in the fluency judgment.

Table 11 *Durations of* a*coustically disfluent recordings (in ms)*

speaker	prompt	session	closure	release	vowel
1	ɑ	2	233.9	NA	538.5
2	sə	1	NA	170.5	195.5
5	kɑ	1	536.4	72.3	315.6
5	ki	1	170.5	114.7	186.7
5	kə	1	106.1	77.1	202.4
5	kə	1	892.3	190.9	165.1
5	kə	1	1596.4	147.8	298.6
5	tə	1	61.0	32.8	199.8
5	tə	1	87.1	273.3	427.4
5	ə	1	219.3	NA	388.2
5	ki	2	216.0	NA	NA
5	pi	2	178.4	26.4	394.7
6	kɑ	1	118.6	86.2	304.9
7	ɑ	1	148.2	NA	326.1
7	pə	1	126.4	60.1	116.0
8	ki	1	159.7	407.3	NA
10	kɑ	1	95.4	54	217.5

What should be noted is that the fluency judgement was performed by a single person, the experimenter. The experimenter did not have prior experience in judging disfluent speech and had to instead rely on her world knowledge. To prevent that exposure to disfluent speech impacted the standards used for the judgement, recordings judged early were reassessed at the end of the process.

The more the experimenter was exposed to potentially disfluent recordings, the more informed was the decision making. This may have influenced judgements made later in the decision process. To avoid potential judgement bias, recordings that were judged in the beginning of the judgement process were reassessed for fluency.

Overall, the perceptual judgement returned 17 perceptually disfluent recordings (see Table 11). For recordings where C = /k, p, s, t/ 14 recordings were judged to be disfluent. These 17 disfluent recordings were excluded from following acoustic analysis (see section 3) and revisited later (see section 5.3).

2.5.2.2 Fluency Judgement of Articulatory Data

Two steps were undertaken to divide the articulatory material into fluent and disfluent and to subset the perceptually disfluent tokens from the overall material: first, the primary researcher undertook a highly cautious / inclusive preselection of potentially disfluent recordings based on the acoustic and articulatory data, and second, a perception study was conducted where five trained linguists were asked to categorise the preselected recordings as fluent and disfluent.

2.5.2.2.1 Pre-Selection

In a first step, the experimenter viewed all the recordings for deviance. Recordings that were different from others were extracted and used as a pre-selection for the subsequent fluency judgement. The criteria for the pre-screening were wide and unspecific which was intended as this step was not meant to form the final categorisation. Instead, it was intended to form a more manageable pre-screened shortlist of recordings for a subsequent fluency judgement (see 2.5.2.2.2). The aim of this first judgement was to decrease the number of recordings that needed to be judged perceptually without excluding potentially disfluent recordings.

All articulatory CV recordings where C = /k/ were investigated visually and perceptually. Each recording was judged in direct comparison to the recordings of the same CV utterance for which acoustic (including sound wave and spectrogram) as well as the articulatory (i.e., ultrasound tongue imaging) information were

consulted. Like the acoustic fluency judgement (see 2.5.2) recordings of each speaker were judged at a time. This approach allowed the experimenter to judge the deviance of recordings against the productions of that same speaker, thereby taking into account the speaker's individual characteristics.

Recordings were considered deviant when they were articulatorily or acoustically different from productions of the same prompt, produced by the same speaker. These recordings were extracted and labelled as deviant. Altogether we extracted 25 CV recordings from the material of all participants who stammer (644 recordings). The extraction process was based on holistic examination of audio and kinetic information from the ultrasound recordings. Both acoustic and articulatory data provided more grounds for deviancy when compared to acoustic judgements alone. This resulted in 25 pre-screened recordings which included overt disfluency, covert disfluency as well as potentially non-categorised behaviour.

2.5.2.2.2 Final Selection

In a second judgement task, the 25 pre-screened deviant recordings were presented to five listeners to judge the recordings on overt fluency.

For the fluency judgement task, listeners were presented with a combination of the preselected deviant recordings and fluent control recordings. For each of the deviant recordings a representative standard fluent recording was chosen as a fluent control recording. The fluent control recordings were defined by contrasting them to recordings of the same CV utterance and produced by the same speaker.

Overall listeners were asked to judge 150 recordings. The material consisted of the 25 recordings that were labelled deviant and their 25 fluent control recordings (totalling 50), which were randomised and presented in three blocks (totalling 150). Materials were randomised for each judge, thereby controlling for order effects.

The CV utterances were presented in a Multiple Forced Choice (MFC) experiment in Praat (Boersma & Weenink, 2015). Listeners were phonetically trained linguists who

were naïve to stammered speech. They did not have any specific training to distinguish disfluent speech from fluent speech but were required to refer to their 'world knowledge' when rating the productions of utterances as either fluent or disfluent. Prior to the MFC, listeners were introduced to the nature of the stimuli they would perceive during the experiment (i.e., V_PCV where V_P = /ə/, C = /k/, V = /ɑ, i, ə /). Participants were informed that the materials were taken from a larger corpus and that they included fluent and disfluent speech.

Participants listened to the recordings one by one. For each recording listeners had to indicate whether they found the recording to be fluent or disfluent (Figure 13).

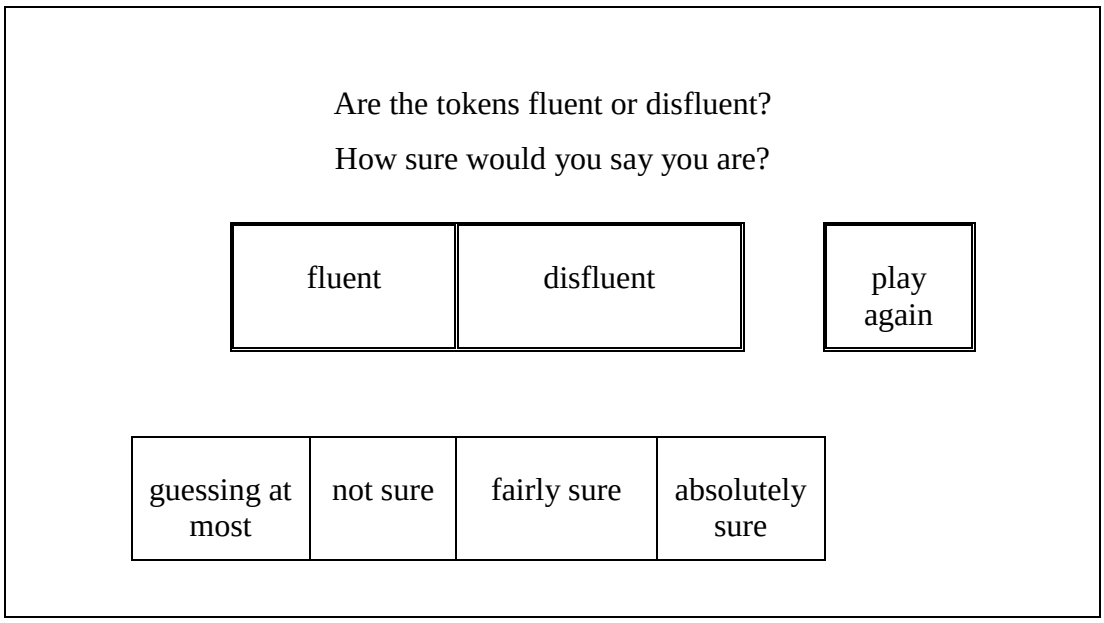


Figure 13 *Layout of the MFC experiment with the binary response question enabled. Listeners are asked whether they perceive the recording as fluent or disfluent.*

Pressing a 'play again' button, recordings could be played up to three times before making a decision. Once recordings were played three times the 'Play again' button disappeared. Following the binary decision about the fluency of the recording (fluent vs. disfluent), a second question was unlocked (changing the button colour from grey to yellow) where listeners were requested to indicate how certain they

were about the binary fluency decision. Using a 4-point scale ranging from 'guessing at most' to 'absolutely sure' each fluency decision could be modified (Figure 14).

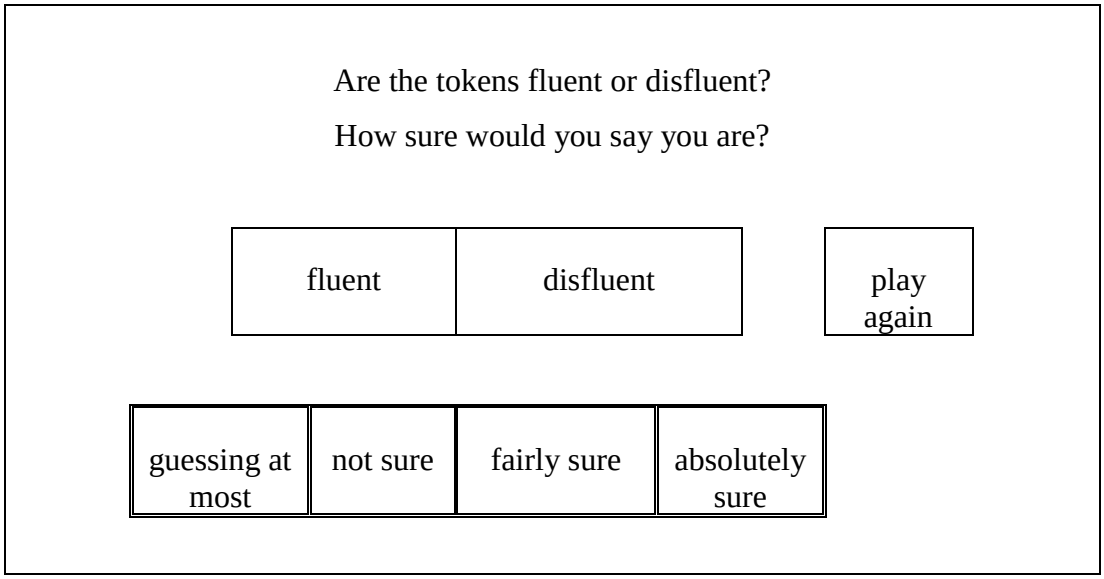


Figure 14 Layout of the MFC experiment *with four goodness categories enabled.*

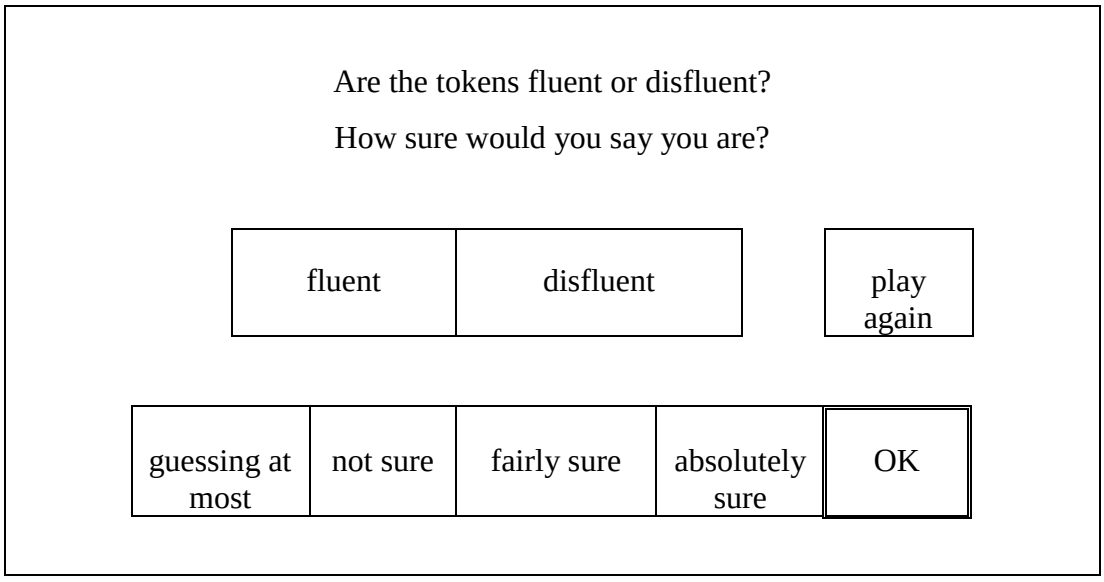


Figure 15 Layout of the MFC experiment *displaying the ok button for final submission of the both the fluency and certainty decisions.*

An 'OK' button was activated after indicating how certain participants were. Pressing the 'OK' button would allow the participant to move on to judge the next

recording. Answers could be revised at any time before pressing the 'OK' button. Both the fluency rating and the indication of certainty had to be provided for each recording before moving on to the next recording by clicking 'OK' (Figure 15). After every 25 recordings, participants were offered a break. At any time during the break participants could indicate that they were ready to continue the task by pressing a mouse button or any key on the keyboard. No recording could be skipped. All five judges completed the experiment.

The objective of the judgement task was to distinguish overtly disfluent recordings from all other recordings. To categorise the recordings, we applied relatively strict criteria: Recordings were labelled disfluent when at least 4 of the 5 listeners agreed on disfluency and more than half (at least 3) of these listeners indicated that they were certain about their decision (3 or 4 points on the 4-point certainty scale). In cases where only few listeners agreed that a recording was disfluent or where listeners were not certain of their judgement, recordings were categorised as fluent and thereby included in the subsequent analysis.

When analysing the ratings from the five listeners, we applied these strict criteria to filter out the clearly disfluent recordings. We found that the criteria needed to be rigorous given that listeners were informed about the nature of the recordings before the judgement task. They were informed that they would be listening to fluent and disfluent recordings, which is highly likely to have influenced their focus of attention and therefore also their judgement leading to an increased proportion of disfluent ratings. The articulatory fluency judgement returned a total of 7 recordings that were categorised as disfluent, leaving. 637 recordings for the analysis.

2.6 Statistical Analysis

Data were analysed using linear mixed effects models (R H Baayen, 2008; R H Baayen, Davidson, & Bates, 2008) using R (version 3.2.5) in R Studio (version 0.99.896) running the lme4 package (version 1.1-12; Bates et al., 2015).

Linear models (Winter, 2013) are useful when modelling a single response variable as a function of multiple predictors / fixed effects including. An error term represents the deviation from the prediction. To account for independence of the data points, we applied mixed effects models (Winter, 2015) where we included individual speaker and recording session as random effects.

All models included the maximal justified random effects structure. We took a forward stepwise approach when adding fixed effect predictors to allow us to explore effects from a theoretical basis. At each step model fit was compared to that of the previous model in order to determine whether the additional predictor improved model fit (i.e., had explanatory value).

In addition to the linear mixed effects models we employed measures to explore the homogeneity and variation for the two speaker groups. The Fligner-Killeen Test of Homogeneity of Variance (Conover, Johnson, & Johnson, 1981) was applied to test the null hypothesis that variances of the two groups are homogenous. The Fligner-Killeen Test is a non-parametric test was employed as it is considered very robust against deviations from normal distribution (Arantes, Eriksson, & Gutzeit, 2017; Jacks & Haley, 2015).

Once the null hypothesis could be rejected, we applied the coefficient of variation to explore the degree of variation by speaker group where larger variation is generally associated with poorer performance (Jäncke, 1994; Olander, Smith, & Zelaznik, 2010).

3 Acoustic Analysis

The following chapter presents the acoustic analysis, including methodology, results and a brief discussion of those results. The methodology section (section 3.1) presents information on data treatment, such as acoustic landmarking (section 3.1.1) and formant extraction (section 3.1.2), as well as the measures applied (section 3.1.3), which comprise measures of segment duration, locus equation and formant slope. Results for these measures (section 3.2) and their implications will be discussed (section 3.3) before we turn to the articulatory analysis (chapter 4).

We present the analysis of acoustic data from 9 people who stammer and 9 control speakers. Data of one speaker (PWS 3) were excluded from analysis due to the strong influence of his second L1. The data from another three participants (PWS 4, 9 and 11) were excluded due to poor quality of the ultrasound image. Additionally, data from control speakers (PNS F and G) were excluded from further analysis.

Within the bigger data set that was recorded we will examine the utterances with onsets /p, t, k, s/ followed by rhymes by /ɑ, i, ə/ that were produced in isolation in the typically voiced condition. After exclusion of disfluent recordings, a total of 627 recordings were investigated for differences in duration as well as formant slope and locus equation.

3.1 Methodology

The acoustic data were stored as independent sound files (in *.wav format) in AAA. All files were exported together with an accompanying text file containing information about the speaker, CV utterance and time / date of the recording.

Recordings were segmented, and the acoustic components (see Table 12) were annotated employing Praat (Boersma & Weenink, 2015). This resulted in four segments. In an initial data screening, we ensured that speakers differentiated between /ə Cɑ/ and /ə Cə/ (see section 2.5.1). We further distinguished acoustically

fluent from disfluent recordings before conducting statistical analysis on the fluent productions of the /ə CV/ utterance (see section 2.5.2).

3.1.1 Acoustic Landmarking

Segmentation and annotation of the sound files were done using Praat. A Praat script (Boersma & Weenink, 2015) was run to read in each sound file and to distinguished audible speech from surrounding and intervening silences. The Praat script inserted boundaries separating sound from silence in each recording. The information about the automatically set boundaries was stored in TextGrid files that were created for each sound file.

Table 12 *Acoustic segmentation*

Annotation	Denotation & Identification
schwa	includes high intensity periodic (and aperiodic) energy before closure
closure	extends from the beginning of the rapid drop in intensity corresponding to the blocked airstream at stop closure, up to the burst
release	aperiodic energy that extends from the burst to the onset of periodic voicing for the CV vowel, incorporating burst energy and aspiration
fricative	aperiodic energy that extends from the end of the periodic energy of the schwa to the onset of periodic voicing for the CV vowel, incorporating burst energy and frication
vowel	high intensity periodic energy following the release or fricative

The automatic segmentation required checking and refinement where we ran a second Praat script. This script allowed for inspection and manual correction of the automatically inserted boundaries. Sound files and the associated TextGrid files

were opened and investigated one by one and the automated boundaries were adjusted if necessary. In places where the speech signal contained several acoustic events that are directly adjacent (e.g., release and vowel in /ə kɑ/), boundaries were inserted manually so that each recording of a CV prompt contained boundaries delimiting each of the up to four acoustic segments (see *Figure 16*). These four segments include the prothetic schwa as the starting point into the actual target syllable and the subsequent target CV syllable consisting of three segments: the acoustic closure, the release into the vowel and the vowel segment.

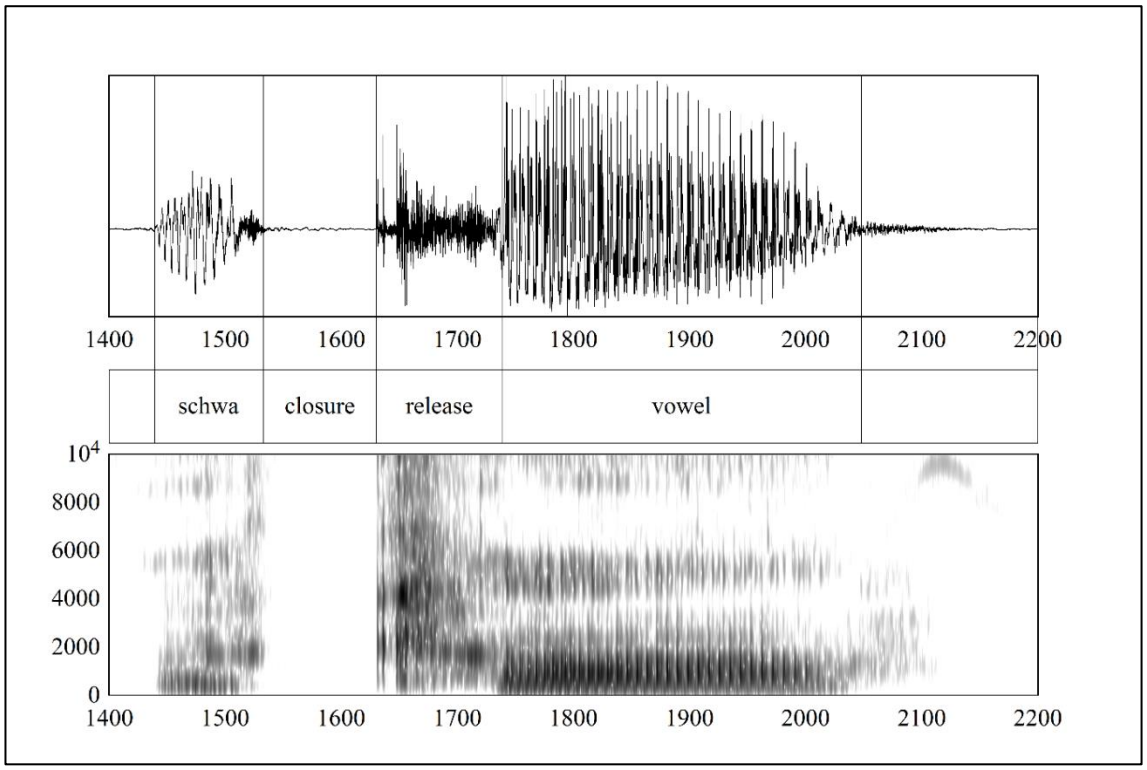


Figure 16 Annotation for CV stimuli where C is a plosive (speaker A, recording 50: /ə kɑ/); three panels showing the sound wave (top), acoustic labels (middle) and spectrogram (bottom); time (in ms) on the x-axis and frequency (in Hz) on the y-axis)

In cases where the consonant of the target syllable was a fricative, no closure phase was marked resulting in three segments (i.e., prothetic schwa, fricative and vowel) that were delimited (see Figure 17). Once the boundaries were inserted and

manually corrected if necessary, a third Praat script assigned labels to the segments of each recording.

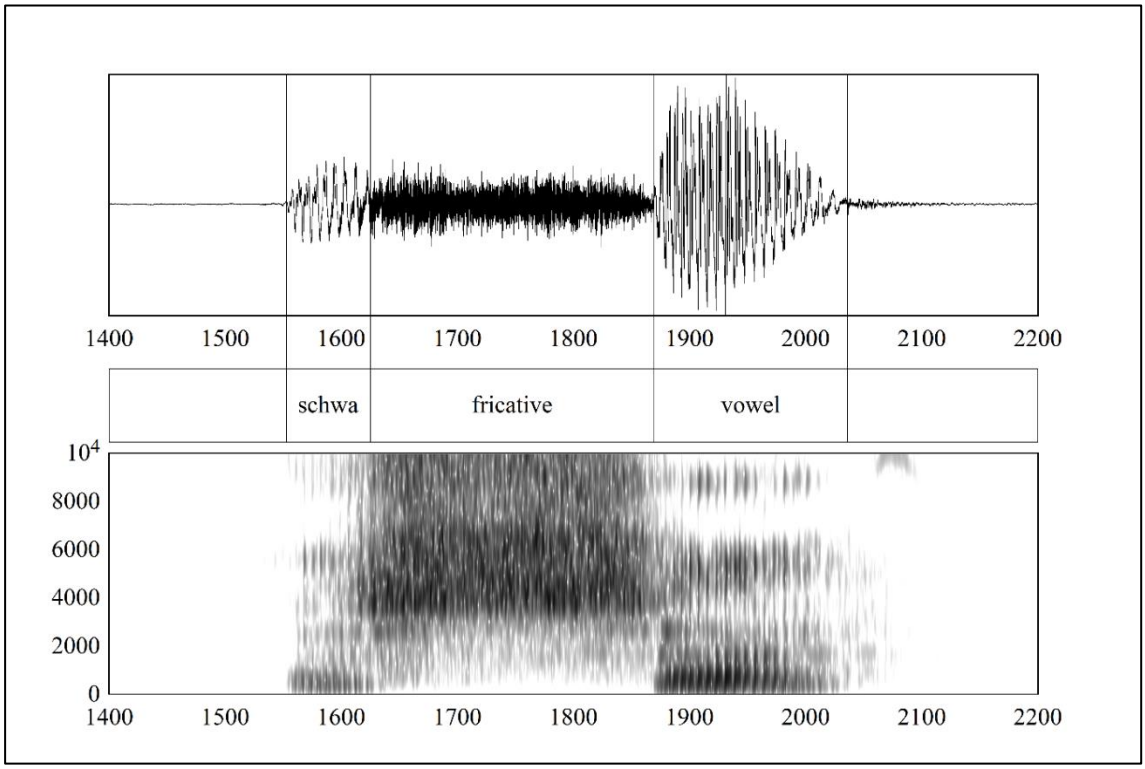


Figure 17 Annotation for CV stimuli where C is a fricative (speaker A, recording 85: /ə sə/); three panels showing the sound wave (top), acoustic labels (middle) and spectrogram (bottom); time (in ms) on the x-axis and frequency (in Hz) on the y-axis)

The boundaries and labels were stored in the TextGrid file accompanying each recording. TextGrid information were then used for both the acoustic as well as the articulatory analysis of the data (Boersma & Weenink, 2015). For the acoustic analysis, boundaries stored in the TextGrid files were employed to investigate segment durations. For articulatory analysis, the newly created TextGrids were imported into the ultrasound software AAA where the boundaries were used for time-aligned articulatory analysis.

3.1.2 Formant Extraction

To explore and compare formant transitions for the two speaker groups of PWS and PNS, we extracted formants from the V_PCV utterances that were categorised as fluent. We adapted an existing Praat script to extract F2 values from the segmented and labelled recordings (see section 3.1.1). The Praat script we adapted is the "msr&check_formants_batch.psc" script from Bert Remijsen dated 20th February 2003 (Remijsen, 2013).

The original script by Bert Remijsen was written to track and extract F1 and F2 values at midpoints of specified segments in a TextGrid. The script contained search window parameters for the formants that would change as a function of gender. During the tracking process, the spectrum for each segment was displayed and available for inspection. Whenever the experimenter felt necessary, tracking values could be adjusted before re-running the script on the displayed segment. The extracted formants were written in a table together with the respective label.

We amended the script to extract F2 values for the labelled segments (i.e., schwa, release, fricative and vowel). Search window parameters for male and female speakers remained unchanged. Formants were extracted at intervals of 10 milliseconds. In addition to formant value and label, several additional variables including speaker group, individual speaker id, recording session, target syllable and time point of the extracted formant, were written to the output table, which was then read into and analysed in RStudio (RStudio Team, 2015).

3.1.3 Acoustic Measures

3.1.3.1 Duration Measures

Each production of a prompt was divided into acoustic segments. Depending on the quality of the consonant, each recording was segmented into three or four constituents (schwa, closure, release and vowel in cases where the consonant was a plosive; schwa, fricative, vowel in cases where the consonant was a fricative).

Especially the closure and the release phases provide information about the transition between the consonant and subsequent vowel. The posture for the prothetic schwa is a means to ensure a stable articulatory starting position across repetitions of prompts (Dayalu et al., 2001). It is, however, not an essential part of the target syllable – especially with the syllable boundary directly following. We therefore decided to not include it in the following analysis, which is targeted specifically at the transition from the consonant onto the following vowel.

3.1.3.2 Locus Equation

Locus equations in contrast to duration measures are a more direct measure of articulation. More precisely, locus equations provide insights into the overall degree of coarticulation as a measure of transition (see 1.5.2.1).

For the present study, locus equations were measured for the different consonant vowel combinations including consonants /k, p, s, t/ and low (/ɑ/), mid (/ə/) and high (/i/) vowels. Regression lines were calculated for each consonantal context by speaker group and severity of stammer to investigate stammering-related differences in the degree of coarticulation as well as potential group differences in even perceptually fluent speech.

3.1.3.3 Formant Slope

Formant slopes were obtained for each V_PCV utterance. We extracted F2 formants at 10 millisecond intervals for each CV sequence in Praat (Boersma & Weenink, 2015). The Praat script was adapted from Remijsen (Remijsen, 2013). Following the formant extraction, we determined two measurement points, the F2 onset (beginning of the formant transition) and the F2 target (end point of the formant transition) for the CV target stimuli (see 1.5.2.2).

For the present study, F2 onset of the formant transition is identified as the first glottal pulse following release of the consonant in the CV utterance (Chang et al., 2002; Robb & Blomgren, 1997). Because the materials included in this current study include voiceless stop consonants, we needed to consider longer aspirations

following stop release (Krull, 1989b; Sussman & Shore, 1996). Determining F2 onset at the first glottal pulse following release, we accounted for transitions already occurring during the aspiration phase. While this approach ensured a consistent point of measure for both plosive and fricative consonants, measures are not directly comparable across consonants due to the typically longer release in fricative consonants. The longer release duration allows for a portion of the formant transition to happen during the fricative. In stop consonants, the shorter release phases do not allow for much formant transition, which is more likely to happen during the subsequent vowel. F2 target is defined as the maximum (for rising F2 slopes with V= /i/) or minimum formant frequency (for falling F2 slopes with V=/ɑ, ə/) within the vowel (Krull, 1987).

Both F2 onset and target are defined via spectro-temporal events which we employed to obtain measures of formant slope duration, formant slope extent and formant transition rate (Tasko & Greilick, 2010) to investigate coarticulatory patterns in the fluent speech of PWS and PNS.

In the present study, slope durations indicate how quickly (starting at the onset) the target within the vowel is reached. The measure of formant slope duration was obtained as the temporal lag between locus and target. Formant slope extent informs about the frequency difference (in Hertz) that needs to be bridged when moving from the formant at onset to the formant at target. The extent of the formant slope was derived from the intensity difference between F2 onset and F2 at target. The third measure was that of formant transition rate, which was obtained by relating the slope extent to the slope duration.

3.2 Results

We analysed the acoustic data from 9 PWS and 9 PNS. The data were balanced by group (PWS = 1260; PNS = 1324), consonant (/k/ = 650, /p/ = 642, /t/ = 645, /s/ = 647), vowel (/ɑ/ = 861, /i/ = 859, /ə/ = 864) and recording session (session 1 = 1302,

session 2 = 1282). Data included 14 disfluent recordings and 2570 fluent recordings for analysis (see Table 13).

Measures of segment duration, as well as formant measures were employed in the investigation of the transition from consonant to subsequent vowel. While measures of duration may indicate differences in overall speech rate, measures of formant trajectories provide more detailed information about the manner with which speakers move between segments of speech.

3.2.1 Segment Duration

Data consisted of CV sequences with a preceding prothetic schwa (V_P). The acoustic signal of each V_PCV utterance was segmented into prothetic schwa, closure, release, and vowel segment. To explore potential acoustic differences in the transition from consonant to vowel we focussed on closure and release / fricative durations for the two speaker groups of PWS and PNS (see Table 13).

Table 13 *Mean and SD durations (in ms) for acoustic closure and release segments by speaker group, consonant, vowel and fluency*

	/k/		/t/		/p/		/s/
	closure	release	closure	release	closure	release	fricative
All fluent	94.44	102.62	100.66	94.71	111.53	81.01	203.89
	(35.65)	(33.16)	(38.96)	(36.75)	(35.08)	(45.66)	(49.12)
PWS	103.40	110.22	108.20	103.70	119.36	89.61	215.61
fluent	(36.04)	(36.05)	(37.99)	(43.18)	(33.87)	(44.56)	(51.43)
/ɑ/	106.28	105.89	108.64	107.97	119.41	89.54	212.83
	(35.63)	(37.75)	(37.89)	(39.66)	(89.54)	(47.65)	(49.81)
/ə/	104.83	109.40	110.52	105.50	121.99	90.59	217.40
	(35.19)	(37.20)	(39.02)	(47.19)	(37.40)	(43.38)	(56.30)
/i/	99.06	115.38	108.64	107.97	119.41	89.54	216.64
	(37.21)	(32.67)	(37.89)	(39.66)	(31.94)	(47.65)	(48.22)
PNS fluent	85.09	94.68	92.98	85.50	103.47	72.14	191.62
	(32.77)	(27.75)	(38.49)	(25.68)	(34.53)	(45.15)	(43.39)
/ɑ/	85.63	90.64	89.96	86.30	99.17	72.63	182.86
	(29.47)	(32.21)	(37.02)	(32.06)	(30.68)	(41.93)	(41.13)
/ə/	87.39	90.69	94.23	84.96	107.42	65.63	197.21
	(29.90)	(24.90)	(38.94)	(23.85)	(32.67)	(19.72)	(45.12)
/i/	82.26	102.67	94.67	85.27	103.90	78.16	194.98
	(38.26)	(23.93)	(39.67)	(20.03)	(39.50)	(62.72)	(42.93)
PWS disfl	588.36	122.30	74.05	153.04	162.00	43.94	170.06
	(573.60)	(42.94)	(18.44)	(170.07)	(50.60)	(21.52)	(NA)

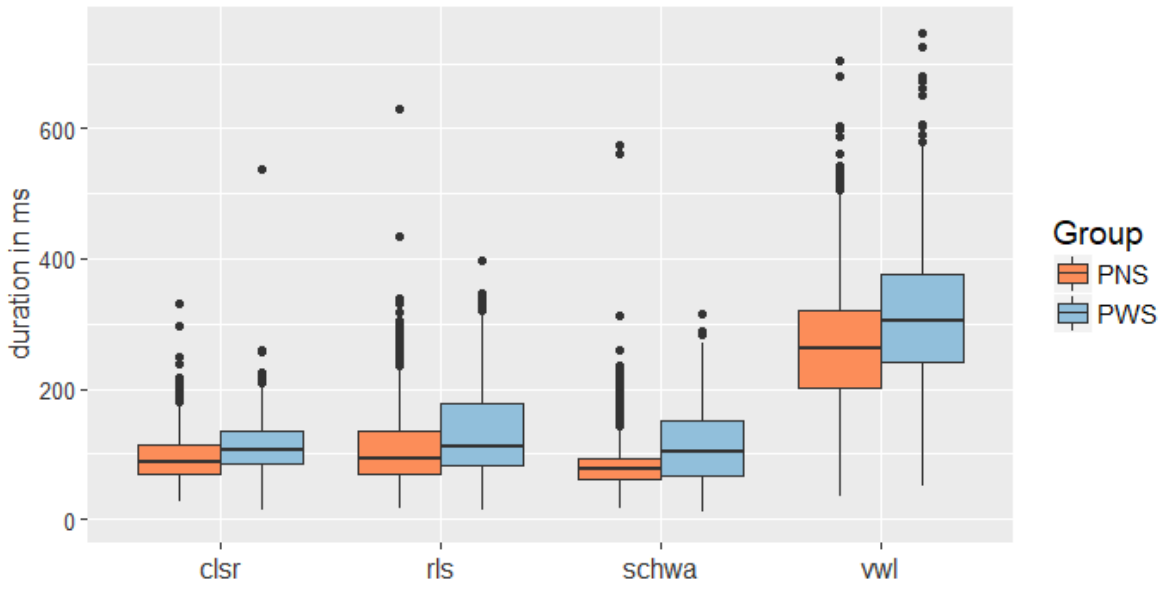


Figure 18 *Acoustic segment durations (in ms) for prothetic schwa, closure, release and subsequent vowel by speaker group*

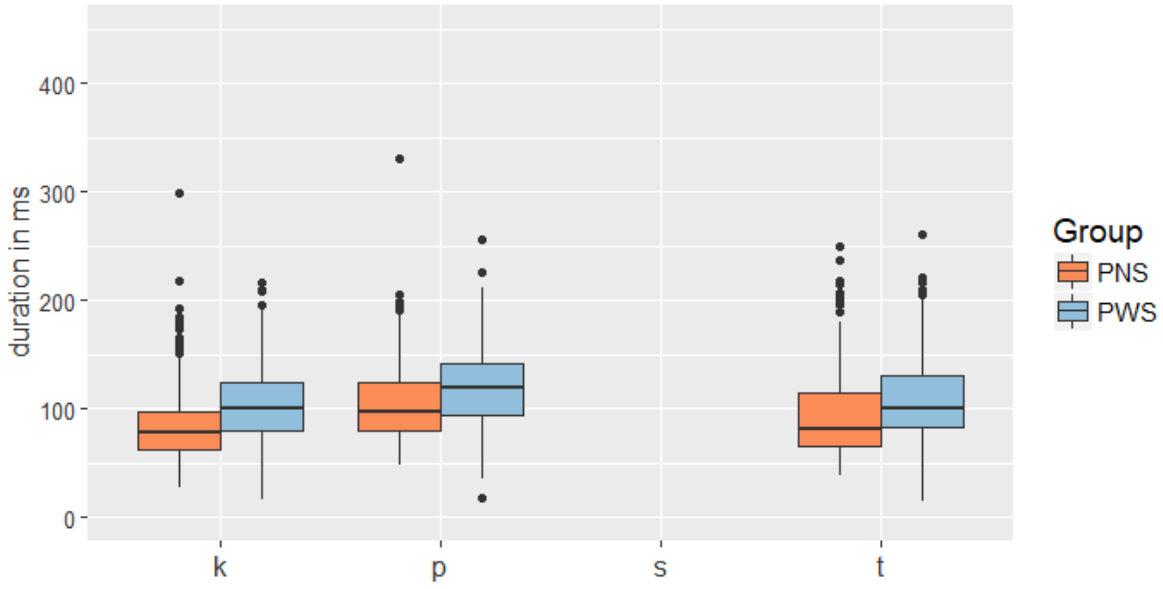


Figure 19 A*coustic closure durations (in ms) by consonant and speaker group*

The segment durations for closure (clsr) are shortest followed by release (rls) durations and meaningfully longer and more variable vowel (vwl) durations. When distinguishing segment durations by speaker group, a tendency for overall longer segment durations can be observed for PWS as compared to PNS. Though this still requires statistical confirmation, descriptively this is a trend that holds across all acoustic segment durations when pooling all consonant and vowel contexts (see Figure 18).

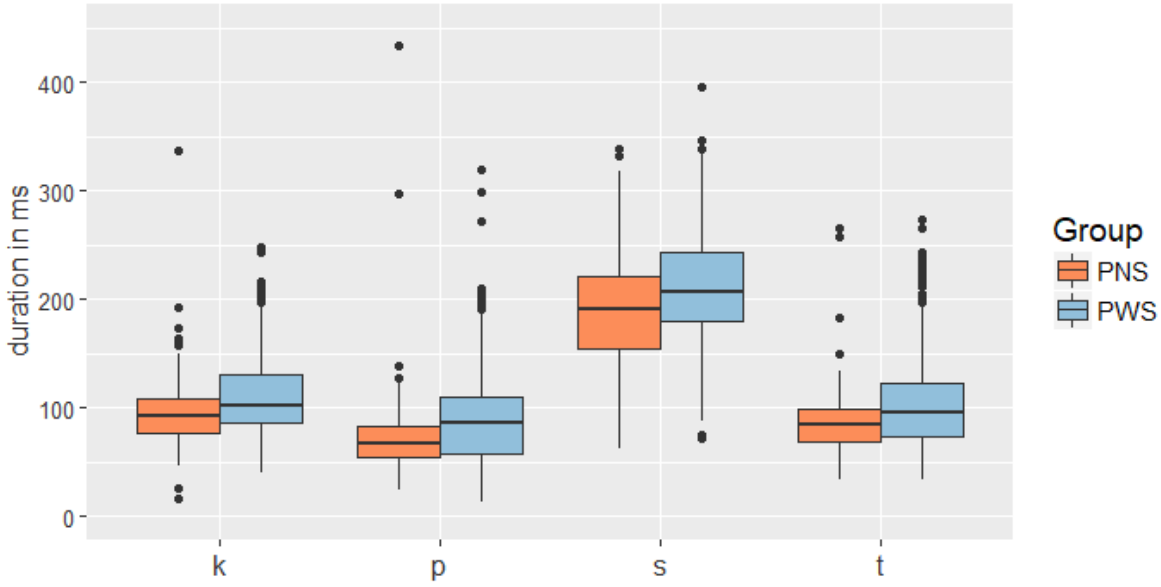


Figure 20 A*coustic release / fricative durations (in ms) by consonant and speaker group*

Regarding the segments that are most relevant to the transition between consonant and vowel, we observe overall longer closure durations (Figure 19) as well as longer release durations (Figure 20) for PWS across the different consonant environments. Standard deviation appears to be high in bilabial environments for PNS, which is likely to be driven by outliers shown in Figure 20. Overall, however, slightly larger variation can be observed for PWS in closure durations (Figure 19) as well as release durations (Figure 20) when compared to PNS.

3.2.1.1 Statistical Analysis for kV

Following the descriptive observations presented in the previous section (see section 3.2.1), we conduct statistical analysis where we treat the durational differences separately for the different consonants and speaker groups. Looking at the consonants one at a time removes the factor of consonant and allows us to better explore differences between the two speaker groups. Analysis was performed on fluent productions exclusively. We first report results for acoustic segment durations by consonant. Data were modelled in milliseconds.

The null model allowed the intercept and slopes to vary by session and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was segment type (i.e., closure, release, and vowel). Including segment type improved model fit ($X^2_{(2)} = 2362.84$, $p < 0.001$). Release durations on average are significantly longer than closure durations ($\beta = 8.18$, $SE(\beta) = 3.21$, $t = 2.55$) and as would be anticipated (Luce & Charles-Luce, 1985) vowel durations were on average longer than closure durations ($\beta = 194.51$, $SE(\beta) = 3.21$, $t = 60.62$).

Subsequently we included vowel (/ɑ/, /ə/, /i/) as a fixed effect. The addition of vowel as a fixed effect improved the model fit ($X^2_{(2)} = 8.82$, $p < 0.05$). The reference level was set at /ɑ/. As anticipated (Browman & Goldstein, 1992b; Koopmans-Van Beinum, 1993), durations for /ə/ were on average significantly shorter than those for /ɑ/ ($\beta = -31.35$, $SE(\beta) = 5.63$, $t = -5.56$). There was, however, no significant difference between /ɑ/ and /i/ ($\beta = -5.25$, $SE(\beta) = 6.08$, $t = -0.86$).

We next included the interaction between segment type and vowel, which further improved model fit ($X^2_{(4)} = 240.94$, $p < 0.001$). Examination of the model indicated that the duration difference between closure and release was increased for /i/ compared to /ɑ/ ($\beta = 16.08$, $SE(\beta) = 7.39$, $t = 2.18$). No effect was found for the closure/release difference for /ə/ compared to /ɑ/ ($\beta = 1.70$, $SE(\beta) = 7.38$, $t = 0.23$). The closure/vowel difference was reduced for /i/ ($\beta = -16.27$, $SE(\beta) = 7.39$, $t = -2.20$).

A larger effect reducing the closure/vowel difference could be observed for /ə/ as compared to /ɑ/ (β = -96.22, SE(β) = 7.38, t = -13.04).

We now further included the main variable of interest, speaker group (PNS, PWS) as fixed effect. The inclusion of speaker group as fixed effect further improved model fit ($X^2_{(1)}$ = 9.98, p < 0.001). Inspection of the model indicated that on average durations were longer for PWS compared to PNS (β = 29.51, SE(β) = 8.85, t = 3.34).

Finally, we added the three-way interaction between segment type, vowel and speaker group. The inclusion of that interaction improved model fit ($X^2_{(8)}$ = 27.17, p < 0.001). Examination of this best fit model indicated that there was no interaction between segment type and speaker group for neither release (β = -5.39, SE(β) = 10.41, t = -0.52) nor vowel durations (β = 15.60, SE(β) = 10.39, t = 1.50).

The three-way interaction between segment type, vowel and group indicated that the closure/release difference that is increased for /i/ compared to /ɑ/ does not differ between groups (β = 1.30, SE(β) = 14.70, t = 0.09). Likewise, is the closure/vowel effect that is reduced for /i/ when compared to /ɑ/ not significantly different between groups (β = 10.21, SE(β) = 14.70, t = 0.70). The same is true for /ə/ where neither the closure release difference (β = 6.66, SE(β) = 14.68, t = 0.45) nor the closure vowel difference (β = 17.17, SE(β) = 14.67, t = 1.17) differs between groups. For full details of this model, please see Table 14.

Table 14 *Model coefficients (in ms) for acoustic segment duration for CV (C = /k/)*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	84.48	8.92	9.47	Speaker Intercept	464.70
Variable (reference level = closure)				- /i/ vs. /ɑ/	25.37
- release : closure	5.01	7.44	0.67	- /ə/ vs. /ɑ/	289.10
- vowel : closure	224.11	7.44	30.11	Session Intercept	0.056
Vowel (reference level = /ɑ/)				- /i/ vs. /ɑ/	49.64
- /i/ v /ɑ/	-2.85	9.10	-0.31	- /ə/ vs. /ɑ/	12.74
- /ə/ v /ɑ/	2.36	9.70	0.24	Residual	2881.0
Group (reference level = PNS)	21.95	12.56	1.75		
Interaction Variable : Vowel					
- release : /i/	15.40	10.50	1.47		
- vowel : /i/	-21.52	10.53	-2.04		
- release : /ə/	-1.71	10.50	-0.16		
- vowel : /ə/	-104.97	10.50	-10.00		
Interaction Variable : Group					
- release : PWS	-5.39	10.41	-0.52		
- vowel : PWS	15.60	10.39	1.50		
Interaction Vowel : Group					
- /i/ : PWS	-4.39	10.66	-0.41		
- /ə/ : PWS	4.11	13.14	0.31		
Interaction Variable : Vowel : Group					
- release : /i/ : PWS	1.30	14.70	0.09		
- vowel : /i/ : PWS	10.21	14.70	0.70		
- release : /ə/ : PWS	6.66	14.68	0.45		
- vowel : /ə/ : PWS	17.17	14.67	1.17		

3.2.1.2 Statistical Analysis for tV

The null model allowed the intercept and slopes to vary by session and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was segment type (i.e., vowel, closure, release, and vowel). Including segment type improved model fit ($X^2_{(2)} = 2175.99$, $p < 0.001$). Release durations showed a tendency to be shorter than closure durations (β = -5.85, SE(β) = 3.48, t = -1.68) and as would be anticipated vowel durations were on average longer than closure durations (β = 189.06, SE(β) = 3.48, t = 54.30).

Subsequently we included vowel (/ɑ/, /ə/, /i/) as a fixed effect. The addition of vowel as a fixed effect improved the model fit ($X^2_{(2)} = 95.34$, $p < 0.001$). The reference level was set at /ɑ/. As anticipated, durations for /ə/ were on average significantly shorter than those for /ɑ/ (β = -31.88, SE(β) = 3.36, t = -9.39). Durations for /i/ were also significantly shorter compared to /ɑ/ (β = -7.05, SE(β) = 3.41, t = -2.07).

We next included the interaction between segment type and vowel, which further improved model fit ($X^2_{(4)} = 251.57$, $p < 0.001$). Examination of the model indicated that the duration difference between closure and release was affected for neither /i/ compared to /ɑ/ (β = 0.60, SE(β) = 7.82, t = 0.08) nor for /ə/ compared to /ɑ/ (β = -1.28, SE(β) = 7.79, t = -0.16). The closure/vowel difference was reduced for /i/ (β = -34.59, SE(β) = 7.82, t = -4.42) with an even larger effect reducing the closure/vowel difference for /ə/ (β = -108.66, SE(β) = 7.79, t = -13.94).

We now further included the main variable of interest, speaker group (PNS, PWS) as fixed effect. The inclusion of speaker group as fixed effect further improved model fit ($X^2_{(1)} = 8.65$, $p < 0.01$). Inspection of the model indicated that on average durations were longer for PWS compared to PNS (β = 28.56, SE(β) = 8.98, t = 3.18).

Table 15 *Model coefficients (in ms) for acoustic segment duration for CV (C = /t/)*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	90.23	9.31	9.69	Speaker Intercept	331.49
Variable (reference level = closure)				Session Intercept	37.39
- release : closure	-3.66	7.82	-0.47	Residual	3248.79
- vowel : closure	219.63	7.82	28.10		
Vowel (reference level = /ɑ/)					
- /i/ v /ɑ/	4.87	7.78	0.63		
- /ə/ v /ɑ/	4.24	7.76	0.55		
Group (reference level = PNS)	14.67	11.59	1.27		
Interaction Variable : Vowel					
- release : /i/	-5.74	11.00	-0.52		
- vowel : /i/	-43.26	11.00	-3.93		
- release : /ə/	-5.61	10.98	-0.51		
- vowel : /ə/	-105.86	10.98	-9.65		
Interaction Variable : Group					
- release : PWS	-3.70	10.95	-0.34		
- vowel : PWS	34.26	10.95	3.13		
Interaction Vowel : Group					
- /i/ : PWS	-1.18	10.95	-0.11		
- /ə/ : PWS	1.09	10.91	0.10		
Interaction Variable : Vowel : Group					
- release : /i/ : PWS	12.43	15.47	0.80		
- vowel : /i/ : PWS	17.80	15.47	1.15		
- release : /ə/ : PWS	8.41	15.41	0.55		
- vowel : /ə/ : PWS	-5.19	15.41	-0.34		

Finally, we added the three-way interaction between segment type, vowel and speaker group. The inclusion of that interaction further improved model fit ($X^2_{(8)}$ = 50.89, p < 0.001). Examination of this best fit model indicated that there was no interaction between segment type and speaker group for release (β = -3.749, SE(β) = 10.550, t = -0.355). The vowel/closure effect in comparison was increased for PWS (β = 34.212, SE(β) = 10.550, t = 3.243) segment durations.

The three-way interaction between segment type, vowel and group indicated that the closure/release difference that did not differ between /i/ and /ɑ/ does also not differ between groups (β = -1.18, SE(β) = 10.95, t = -0.11). Likewise, is the closure/vowel effect that is reduced for /i/ when compared to /ɑ/ comparable for both groups (β = 12.43, SE(β) = 15.47, t = 0.80). For full details of this model please see Table 15.

3.2.1.3 Statistical Analysis for pV

The null model allowed the intercept and slopes to vary by session and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was segment type (i.e., vowel, closure, release, and vowel). Including segment type improved model fit ($X^2_{(2)}$ = 2215.86, p < 0.001). Release durations were on average shorter than closure durations (β = -30.50, SE(β) = 3.47, t = -8.80) and as would be anticipated vowel durations were on average longer than closure durations (β = 177.70, SE(β) = 3.47, t = 51.23).

Subsequently we included vowel (/ɑ/, /ə/, /i/) as a fixed effect. The addition of vowel as a fixed effect improved the model fit ($X^2_{(2)}$ = 113.41, p < 0.001). The reference level was set at /ɑ/. Durations for /ə/ were on average significantly shorter than those for /ɑ/ (β = -33.14, SE(β) = 3.36, t = -9.85). There was no significant difference between /ɑ/ and /i/ (β = -3.78, SE(β) = 3.38, t = -1.12).

We next included the interaction between segment type and vowel, which further improved model fit ($X^2_{(4)}$ = 265.21, p < 0.001). Examination of the model indicated that the there was no effect on the duration difference between closure and release

for /i/ and /ɑ/ (β = -0.69, SE(β) = 7.72, t = -0.09) with also no effect on the closure/release difference for /ə/ (β = -9.28, SE(β) = 7.68, t = -1.21). The closure/vowel difference was reduced for /i/ (β = -19.41, SE(β) = 7.72, t = -2.51) with an even larger effect reducing the closure/vowel difference for /ə/ (β = -109.34, SE(β) = 7.68, t = -14.23).

We now further included the main variable of interest, speaker group (PNS, PWS) as fixed effect. The inclusion of speaker group as fixed effect further improved model fit ($X^2_{(1)} = 6.79$, $p < 0.01$). Inspection of the model indicated that on average durations were longer for PWS compared to PNS (β = 24.64, SE(β) = 0.02, t = 2.73).

Finally, we added the three-way interaction between segment type, vowel and speaker group. The inclusion of that interaction further improved model fit ($X^2_{(8)} = 23.20$, $p < 0.01$). Examination of this best fit model indicated that there was no interaction between segment type and speaker group for release (β = -1.18, SE(β) = 10.87, t = -0.11). The closure/vowel difference, however, was increased for PWS (β = 25.48, SE(β) = 10.87, t = 2.35) segment durations.

The three-way interaction between segment type, vowel and group indicated that the closure/release difference that had been decreased in /i/ compared to /ɑ/ does not differ between groups (β = -2.96, SE(β) = 15.38, t = -0.19) compared to PNS. Likewise, is the closure/vowel effect that is reduced for /i/ comparable for both groups (β = 4.60, SE(β) = 15.39, t = 0.30). The closure/vowel difference for /ə/ also does not differ between groups (β = 11.57, SE(β) = 15.31, t = 0.76) compared to PNS. Likewise, is the closure/vowel effect for /ə/ comparable for both groups (β = -5.33, SE(β) = 15.31, t = -0.35). For full details of this model please see Table 16.

Table 16 *Model coefficients (in ms) for acoustic segment duration for CV (C = /p/)*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	98.94	8.90	11.11	Speaker Intercept	336.35
Variable (reference level = closure)				Session Intercept	24.58
- release : closure	-26.55	7.67	-3.46	Residual	3121.25
- vowel : closure	208.27	7.67	27.14		
Vowel (reference level = /ɑ/)					
- /i/ v /ɑ/	4.69	7.71	0.61		
- /ə/ v /ɑ/	8.40	7.71	1.09		
Group (reference level = PNS)	17.97	11.58	1.55		
Interaction Variable : Vowel					
- release : /i/	0.81	10.90	0.07		
- vowel : /i/	-21.98	10. 92	-2.01		
- release : /ə/	-15.24	10.90	-1.40		
- vowel : /ə/	-107.01	10. 90	-9.81		
Interaction Variable : Group					
- release : PWS	-1.18	10.87	-0.11		
- vowel : PWS	25.48	10. 87	2.35		
Interaction Vowel : Group					
- /i/ : PWS	-3.27	10.88	-0.30		
- /ə/ : PWS	-3.64	10.83	0.34		
Interaction Variable : Vowel : Group					
- release : /i/ : PWS	-2.96	15.38	-0.19		
- vowel : /i/ : PWS	4.60	15.39	0.30		
- release : /ə/ : PWS	11.57	15.31	0.76		
- vowel : /ə/ : PWS	-5.33	15.31	-0.35		

3.2.1.4 Statistical Analysis for sV

The null model allowed the intercept and slopes to vary by session and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was segment type (i.e., schwa, release, and vowel). Including segment type improved model fit ($X^2_{(1)} = 531.20$, $p < 0.001$). As would be anticipated vowel durations were on average longer than release durations ($\beta = 95.59$, $SE(\beta) = 3.72$, $t = 25.72$).

Subsequently we included vowel (/ɑ/, /ə/, /i/) as a fixed effect. The addition of vowel as a fixed effect improved the model fit ($X^2_{(2)} = 9.02$, $p < 0.05$). The reference level was set at /ɑ/. As anticipated, durations for /ə/ were on average significantly shorter than those for /ɑ/ ($\beta = -48.27$, $SE(\beta) = 7.89$, $t = -6.12$). Durations for /i/ were on average also shorter compared to /ɑ/ ($\beta = -13.78$, $SE(\beta) = 4.74$, $t = -2.91$).

We next included the interaction between segment type and vowel, which further improved model fit ($X^2_{(2)} = 180.92$, $p < 0.001$). Examination of the model indicated that the duration difference between release and vowel was reduced for /i/ compared to /ɑ/ ($\beta = -41.67$, $SE(\beta) = 8.43$, $t = -4.94$) with an even larger effect reducing the release/vowel difference for /ə/ ($\beta = -116.90$, $SE(\beta) = 8.49$, $t = -13.76$).

We now further included the main variable of interest, speaker group (PNS, PWS) as fixed effect. The inclusion of speaker group as fixed effect further improved model fit ($X^2_{(1)} = 6.86$, $p < 0.01$). Inspection of the model indicated that on average durations were longer for PWS compared to PNS ($\beta = 37.78$, $SE(\beta) = 12.80$, $t = 2.95$).

Finally, we added the three-way interaction between segment type, vowel and speaker group. The inclusion of that interaction further improved model fit ($X^2_{(5)} = 20.18$, $p < 0.01$). Examination of this best fit model indicated that there was no interaction between segment type and speaker group for vowel segment durations ($\beta = -8.06$, $SE(\beta) = 11.84$, $t = -068$).

The three-way interaction between segment type, vowel and group indicated that the vowel/release difference that had been decreased in /i/ compared to /ɑ/ is

further reduced in PWS compared to PNS (β = 48.54, SE(β) = 16.75, t = 2.90) compared to PNS. Likewise, is the vowel/release effect that is reduced for /ə/ further reduced for PWS compared to PNS (β = 36.53, SE(β) = 16.88, t = 2.16). For full details of this model please see Table 17.

Table 17 *Model coefficients (in ms) for acoustic segment duration for CV (C = /s/)*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	181.34	13.71	13.23	Speaker Intercept	1187.00
Variable (reference = fricative)				- /i/ vs. /ɑ/	16.61
- vowel : fricative	151.86	8.46	17.96	- /ə/ vs. /ɑ/	828.5
Vowel (reference level = /ɑ/)				Session Intercept	39.35
- /i/ v /ɑ/	12.40	8.65	1.43	- /i/ vs. /ɑ/	3.53
- /ə/ v /ɑ/	16.38	12.89	1.27	- /ə/ vs. /ɑ/	0.01
Group (reference level = PNS)	32.14	18.23	1.76	Residual	3790.00
Interaction Variable : Vowel					
- vowel : /i/	-66.30	11.95	-5.55		
- vowel : /ə/	-135.66	12.09	-11.22		
Interaction Variable : Group					
- vowel : PWS	-8.06	11.84	-0.68		
Interaction Vowel : Group					
- /i/ : PWS	-10.63	11.98	-0.89		
- /ə/ : PWS	-12.45	18.09	-0.69		
Interaction Variable : Vowel : Group					
- vowel : /i/ : PWS	48.54	16.75	2.90		
- vowel : /ə/ : PWS	36.53	16.88	2.16		

3.2.1.5 Variation & Homogeneity

Now that we have explored the durational differences for the segment durations by consonant, where we find PWS to present with overall longer acoustic segment durations, we are interested to see how homogenous / variable the two speaker groups are.

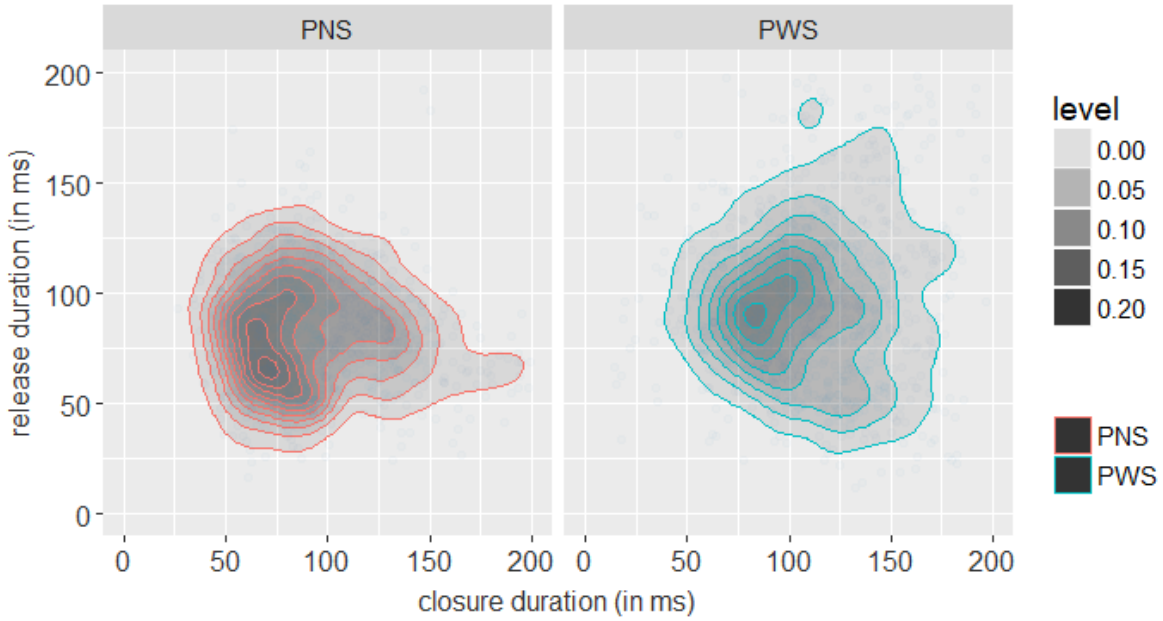


Figure 21 *Density plot for closure durations (x-axis) by release durations (y-axis) by speaker group*

In a first step we have explored the homogeneity of variance of the segment durations for the two speaker groups. We applied the Fligner-Killeen Test of Homogeneity of Variance to prompts where C = /k, p, t/. Speaker group was entered as independent variable with segment duration as dependent variable. The test was run separately by segment type (closure, release, and vowel). The Fligner-Killeen Test of Homogeneity of Variance returned significant differences for the homogeneity of the two speaker groups for closure durations ($X^2_{(1)} = 10.22$, $p < 0.01$) as well as release ($X^2_{(1)} = 81.68$, $p < 0.001$) and vowel duration ($X^2_{(1)} = 10.54$, p

< 0.01; see Table 18 and Figure 21). We can therefore reject the null hypothesis that variances of the durations of the two groups are homogenous.

Next, we explored the variability for the two speaker groups. We employed the measure of coefficient of variation to evaluate variation while normalising the different mean values. We applied the coefficient of variation to the segment durations (including closure, release and vowel) with speaker groups pooled (overall) as well as to the durations of acoustic closure, release and vowel phases for the two speaker groups separately. The coefficient of variation is highest for closure (52.56%) and release durations (43.21%) and lowest for vowel durations (34.41%).

By speaker group, the coefficient of variation returned differing variances for closure durations (PNS: 38.48%, PWS: 58.51%). In contrast, the coefficient of variation for release durations (PNS: 41.82%, PWS: 42.21%) as well as for vowel durations (PNS: 33.91%, PWS: 33.07%) returned a comparable variance for the two speaker groups (see Table 18 and Figure 21).

Table 18 *Coefficient of variation and homogeneity of variance for acoustic segment durations by group and segment (closure, release and vowel) for C = /k, p, t/*

		homogeneity of variance	coefficient of variation (in %)	
speaker group		PWS vs. PNS	PNS	PWS
segment type	closure	$p < 0.01$	38.48	58.51
	release	$p < 0.001$	41.82	42.21
	vowel	$p < 0.001$	33.91	33.07

3.2.2 Locus Equation

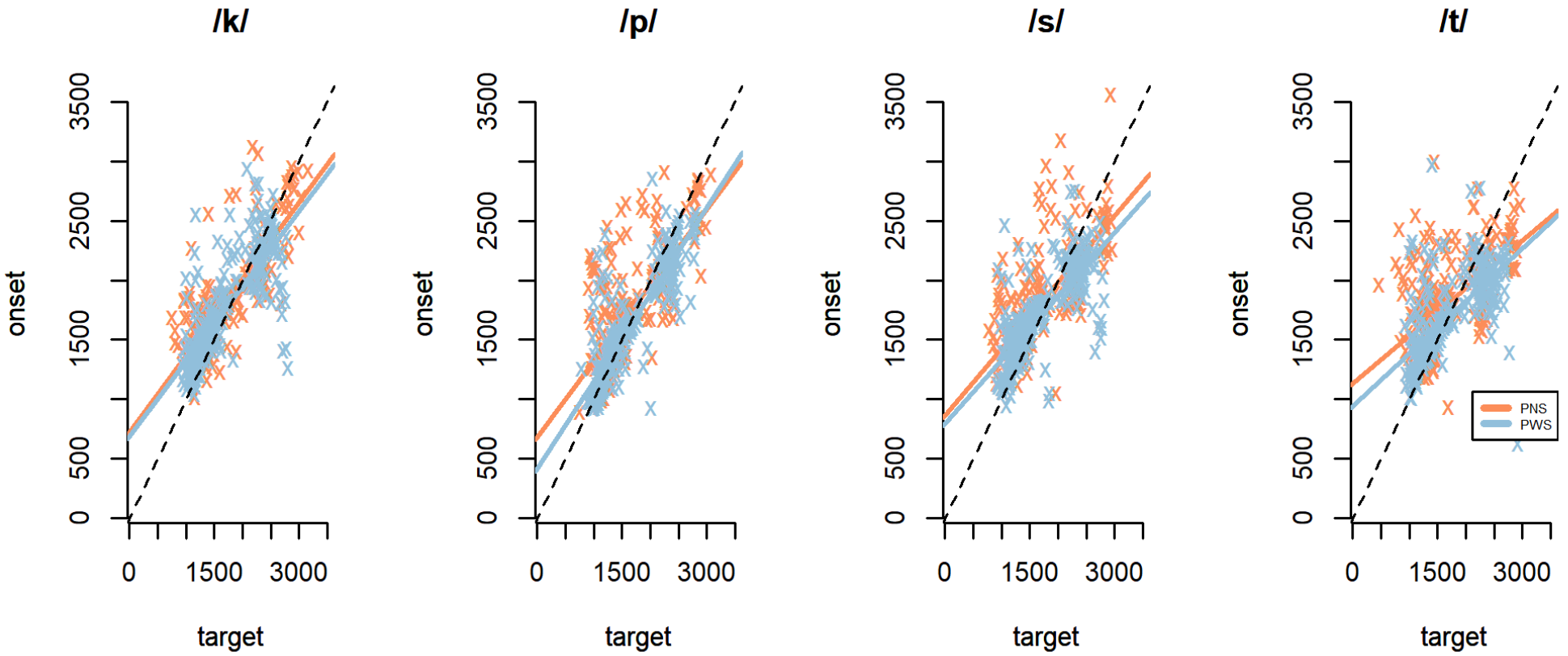


Figure 22 *Locus equations by speaker group comparing PWS and PNS across consonants (/k, p, s, t/)*

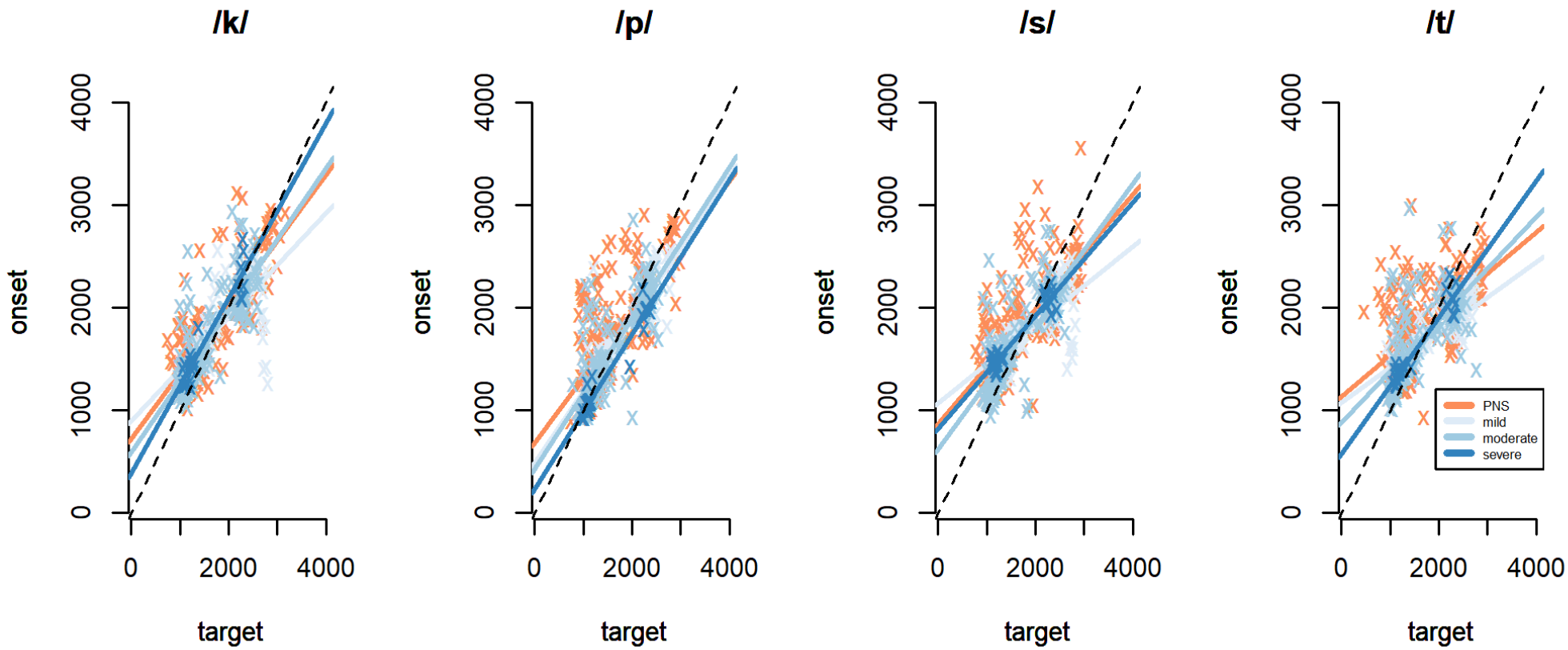


Figure 23 *Locus equations by severity of stammer comparing PWS (by severity of stammer) and PNS across consonants (/k, p, s, t/)*

Table 19 *Locus equations for PWS and PNS across consonant (/k, p, s, t/)*

	PNS		PWS	
consonant	intercept (in Hz)	slope	intercept (in Hz)	slope
/k/	725.69	0.64	686.26	0.63
/p/	684.81	0.64	418.98	0.73
/s/	865.06	0.56	798.27	0.53
/t/	1131.14	0.40	941.83	0.44

Table 20 *Locus equation intercept and slope for consonant (/k, p, s, t/) by speaker group and severity of stammer*

	/p/		/k/		/t/		/s/	
	intercept	slope	intercept	slope	intercept	slope	intercept	slope
PNS	684.81	0.63	725.69	0.64	1131.14	0.40	865.06	0.55
mild	490.26	0.69	896.244.	0.51	1076.36	0.34	1068.99	0.38
moderate	433.16	0.73	595.59	0.68	879.95	0.50	621.97	0.64
severe	236.87	0.75	385.74	0.85	577.36	0.66	823.64	0.55

Steeper slopes reflect a larger degree in coarticulation. When comparing PWS (in blue) to PNS (in orange; see Figure 22) the locus equation regression lines for the two speaker groups overall do not appear to differ between groups (see Table 19) or with severity of the stammer (see Table 20 and Figure 23). Statistically, however, the larger sample size is taken into account revealing significant differences between groups (see Table 21 and Table 22).

3.2.2.1 Statistical Analysis

We conducted statistical analysis where we modelled the locus equations including slope and intercept. Analysis was performed on perceptually fluent productions

exclusively. Data were modelled in Hertz. All models included the maximal justified random effects structure. We took a forward stepwise approach when adding fixed effect predictors to allow us to explore effects from a theoretical basis. At each step, model fit was compared to that of the previous model in order to determine whether the additional predictor improved model fit (i.e., had explanatory value).

Table 21 *Model coefficients for locus equation slopes for CV where C = /k, p, s, t/*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	737.39	49.41	14.92	Speaker Intercept	10734
Slope	0.65	0.02	32.60	Residual	60401
Consonant (reference = /k/)					
- consonant : /p/	-171.74	47.75	-3.60		
- consonant : /t/	330.52	47.95	6.89		
- consonant : /s/	115.78	48.54	2.39		
Group (reference = PNS)	-106.02	49.97	-2.12		
Interaction Slope : Consonant					
- slope : /p/	0.06	0.03	2.03		
- slope : /t/	-0.22	0.03	-7.88		
- slope : /s/	-0.08	0.03	-3.05		

The null model allowed the intercept to vary by session and speaker (i.e., as random effect). The first predictor to be included as a fixed effect was consonant (i.e., /k, p, t, s/). Including consonant improved model fit ($X^2_{(3)} = 33.30$, $p < 0.001$). The reference level was set at /k/. Intercept for /p/ was significantly lower ($\beta = -172.01$, $SE(\beta) = 47.75$, $t = -3.60$) while intercepts for /s/ ($\beta = 115.70$, $SE(\beta) = 48.54$, $t = 2.38$) and /t/ ($\beta = 330.62$, $SE(\beta) = 47.94$, $t = 6.90$) were significantly higher when compared to /k/.

Adding the interaction with consonant further improved model fit ($X^2_{(3)}$ = 110.87, p < 0.001). The reference level was set at /k/. The slope for /p/ was significantly steeper (β = 0.06, SE(β) = 0.03, t = 2.03) while slopes for /s/ (β = -0.08, SE(β) = 0.03, t = -3.04) and /t/ (β = -0.22, SE(β) = 0.03, t = -7.88) were significantly flatter when compared to /k/. This relation of steepness can also be observed in Figure 23.

Next, we included group (PWS, PNS) as fixed effect. Including group improved model fit ($X^2_{(1)}$ = 4.47, p < 0.05). The reference level was set at PNS. Intercept for PWS was significantly lower (β = -106.02, SE(β) = 49.97, t = -2.12) when compared to PNS. Adding the interaction with group did not improve model fit ($X^2_{(7)}$ = 11.23, p = 0.129) and neither did adding severity of the stammer as a fixed factor ($X^2_{(3)}$ = 4.73, p = 0.19).

3.2.2.2 Variation and Homogeneity

Following the investigation of locus equation slopes, we investigate the distribution of F2 onset and F2 target values for both speaker groups (see Table 22).

We applied the measure of homogeneity (Fligner-Killeen Test of Homogeneity of Variance) to the F2 values obtained at onset and at target to compare the distribution speaker group. The results indicate a statistically significant difference for the two speaker groups overall for F2 onset ($X^2_{(1)}$ = 15.27, p < 0.001) and F2 target values ($X^2_{(1)}$ = 9.50, p < 0.01). Adding the consonant environment (C=/p, t, k, s/) as independent factor reveals as significant difference in homogeneity of variance for alveolar consonant environments at F2 onset (t: $X^2_{(1)}$ = 4.03, p < 0.05; /s/: $X^2_{(1)}$ = 14.63, p < 0.001) and F2 target values (t: $X^2_{(1)}$ = 10.22, p < 0.01; /s/: $X^2_{(1)}$ = 5.90, p < 0.05).

Table 22 *Variation and Homogeneity for F2 onset and F2 target values for both speaker groups across the different consonant environments*

	onset			target		
	homogeneity of variance	coefficient of variation		homogeneity of variance	coefficient of variation	
	PNS vs. PWS	PNS	PWS	PNS vs. PWS	PNS	PWS
overall	$p < 0.001$	22.02	19.07	$p < 0.05$	23.33	19.89
/k/	$p = 0.06$	19.97	18.49	$p = 0.06$	22.42	21.13
/p/	$p = 0.13$	22.53	23.35	$p = 0.14$	23.09	22.42
/t/	$p < 0.05$	18.94	19.79	$p < 0.05$	23.10	21.32
/s/	$p < 0.001$	20.58	19.07	$p < 0.05$	21.47	19.89

Next, we applied the coefficient of variation to the onset and target values for both PWS and PNS for the different consonant environments (see Table 22 and Figure 24). Overall, the coefficient of variation for PWS is lower when compared to PNS for onset values (PNS: 22.02%; PWS: 19.07%) as well as target values (PNS: 23.33%; PWS: 19.89%). By consonant, onset values do not present with a clear group divide. In contrast, the coefficients of variation for target values are consistently lower for the people who stammer when compared to typical speakers (/k/: PNS = 22.42%, PWS = 21.13%. /p/: PNS = 23.09%, PWS = 22.42%, /t/: PNS = 23.10%, PWS = 21.32%, /s/: PNS = 21.47%, PWS = 19.89%).

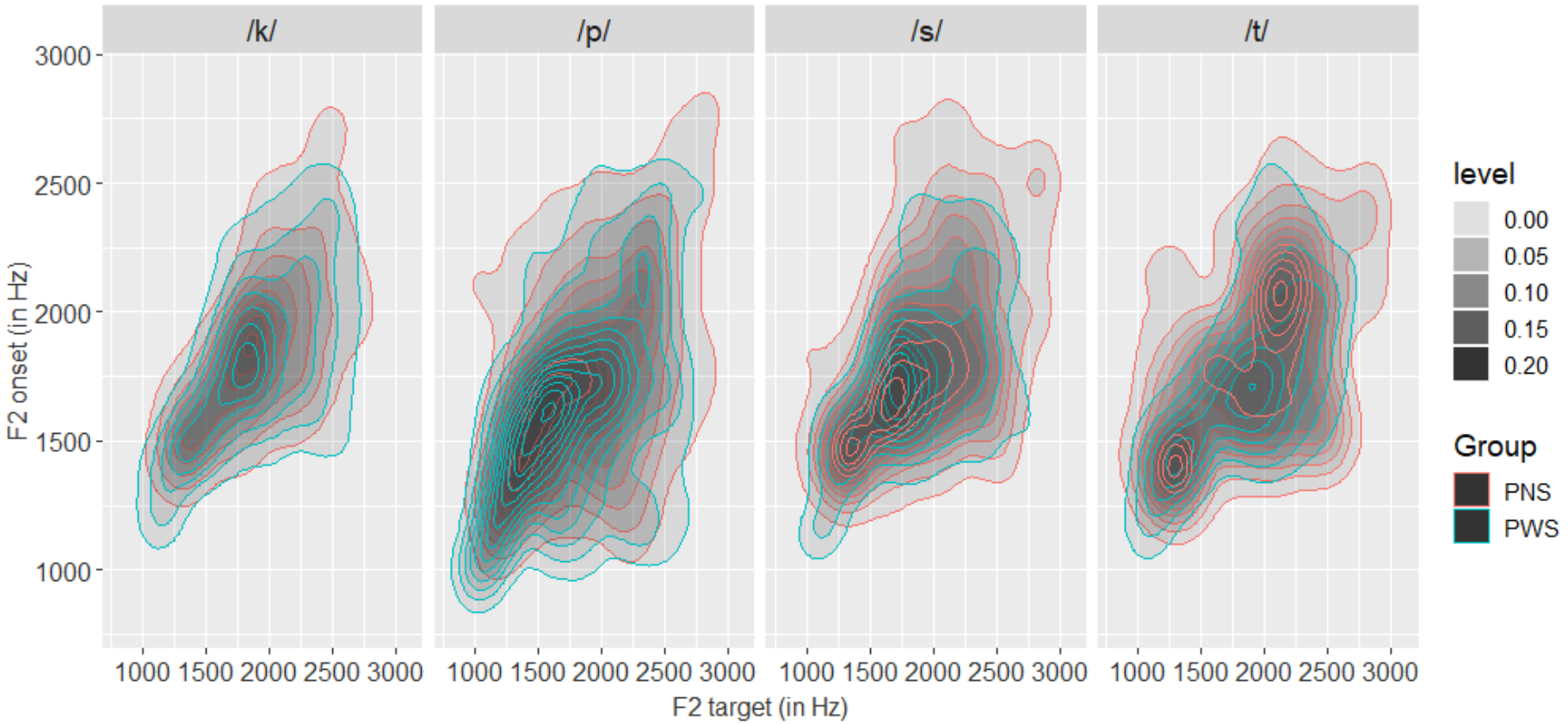


Figure 24 *Density plot displaying F2 onset and F2 target for PWS and PNS across consonants (/k/, /p/, /s/, /t/)*

3.2.3 Formant Slope

We present statistical analysis where we fitted linear mixed effect models. Models were fitted for the F2 slope duration (modelled in milliseconds), the extent of the F2 slope (modelled in Hertz) and the transition rate (see section 3.2.3.3) of the F2 slope (modelled in milliseconds/Hertz). For these three models, we established maximal justified random effects structure before testing the best model fit. To determine the best model fit we took a stepwise forward approach starting with the null model and building up adding fixed effect predictors. The order in which fixed effect predictors are added was theoretically informed. We first report results for the F2 formant slope duration.

3.2.3.1 Duration

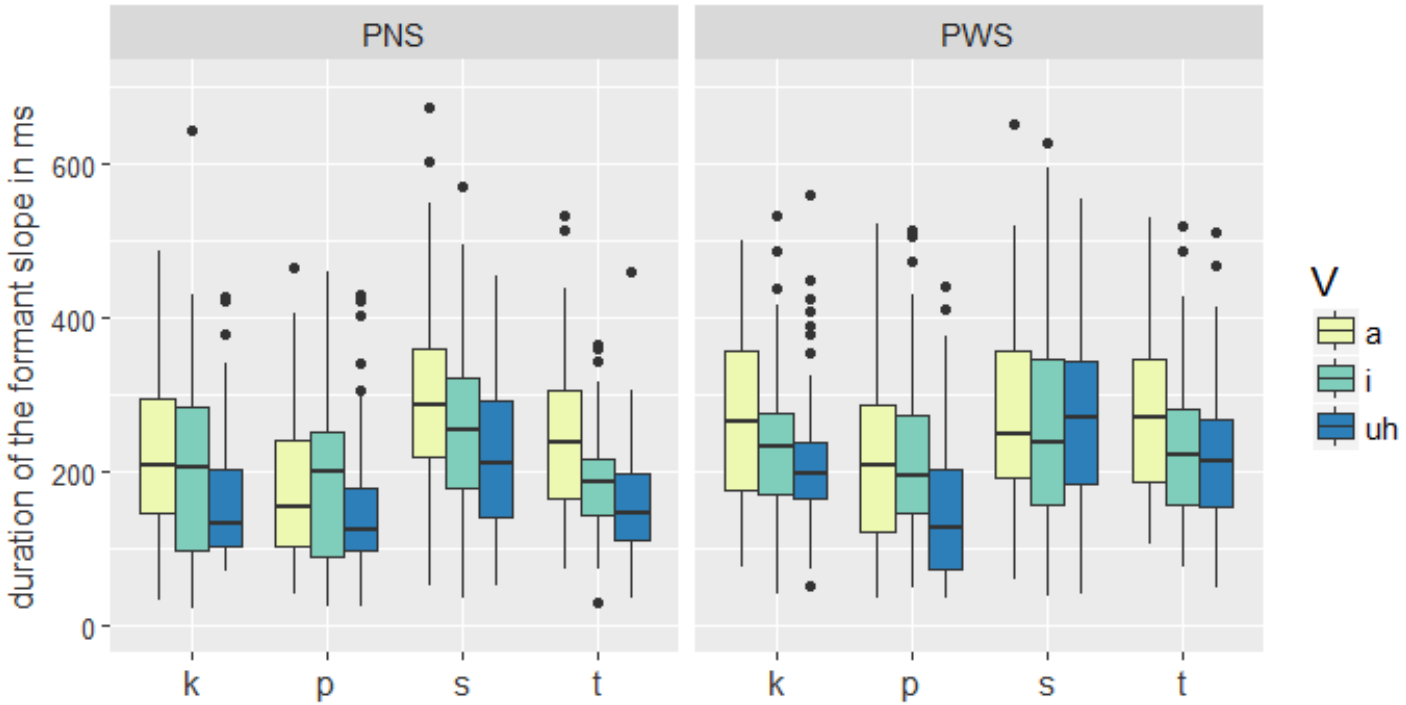


Figure 25 *Formant slope durations for PWS and PNS across consonant (/k, p, s, t/) and vowel (/a, i, ə /) environments*

The null model allowed the intercept and slopes to vary by vowel and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was vowel (i.e., /ɑ, i, ə /). Including vowel improved model fit ($X^2_{(2)}$ = 15.69, p < 0.001). The reference level was set at /ɑ/. Slope durations for /ə/ were on average shorter than slope durations for /ɑ/ (β = -68.42, SE(β) = 14.91, t = -4.59) while slope durations for /i/ (β = -26.45, SE(β) = 14.44, t = -1.83) showed a strong tendency to be shorter, which, however, did not reach significance.

Subsequently we included consonant (/k/, /p/, /s/, /t/) as a fixed effect. The addition of consonant as a fixed effect improved the model fit ($X^2_{(3)}$ = 1642, p < 0.001). The reference level was set at /k/. Durations for /p/ (β = -48.67, SE(β) = 3.23, t = -15.44) and /t/ (β = -9.78, SE(β) = 3.14, t = -3.11) were on average significantly shorter than those for /k/. Slope durations for /s/ in contrast were significantly longer (β = 85.07, SE(β) = 3.15, t = 26.35) when compared to slope duration for /k/ (see Figure 25).

We next included the interaction between vowel and consonant, which further improved model fit ($X^2_{(6)}$ = 97.96, p < 0.001). Examination of the model indicated that the slope duration difference between /ɑ/ and /i/ was significantly decreased for /p/ compared to /k/ (β = 19.71, SE(β) = 7.73, t = 2.55) while the /ɑ/-/i/ difference was not affected in /s/ context (β = 1.4, SE(β) = 7.76, t = 0.18) and increased in /t/ context (β = -47.88, SE(β) = 7.62, t = -6.28). The /ɑ/-/ə/ difference was not affected in /p/ (β = -0.69, SE(β) = 7.60, t = -0.09) or in /s/ context (β = 9.76, SE(β) = 7.9, t = 1.24) while the difference was increased and in /t/ context (β = -23.17, SE(β) = 7.65, t = -3.02) when compared to /k/.

Adding speaker group as a fixed effect did not improve model fit ($X^2_{(1)}$ = 2.97, t = 0.08). For full details of this model please see Table 23.

Table 23 *Model coefficients (in ms) for formant slope duration*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	280.468	12.062	23.252	Speaker Intercept	48.42
Vowel				- /i/ vs. /ɑ/	54.02
(reference level = /ɑ/)					
- /i/ v /ɑ/	-17.587	13.988	-1.257	- /ə/ vs. /ɑ/	51.02
- /ə/ v /ɑ/	-59.030	13.371	-4.415	Residual	87.96
Consonant					
(reference level = /k/)					
- /p/ v /k/	-55.043	5.412	-10.171		
- /s/ v /k/	81.584	5.462	14.938		
- /t/ v /k/	14.416	5.382	2.678		
Interaction					
Vowel : Consonant					
- /i/ : /p/	19.715	7.732	2.550		
- /ə/ : /p/	-0.687	7.600	-0.090		
- /i/ : /s/	1.402	7.758	0.181		
- /ə/ : /s/	9.763	7.900	1.236		
- /i/ : /t/	-47.883	7.624	-6.280		
- /ə/ : /t/	-23.174	7.655	-3.028		

3.2.3.2 Extent

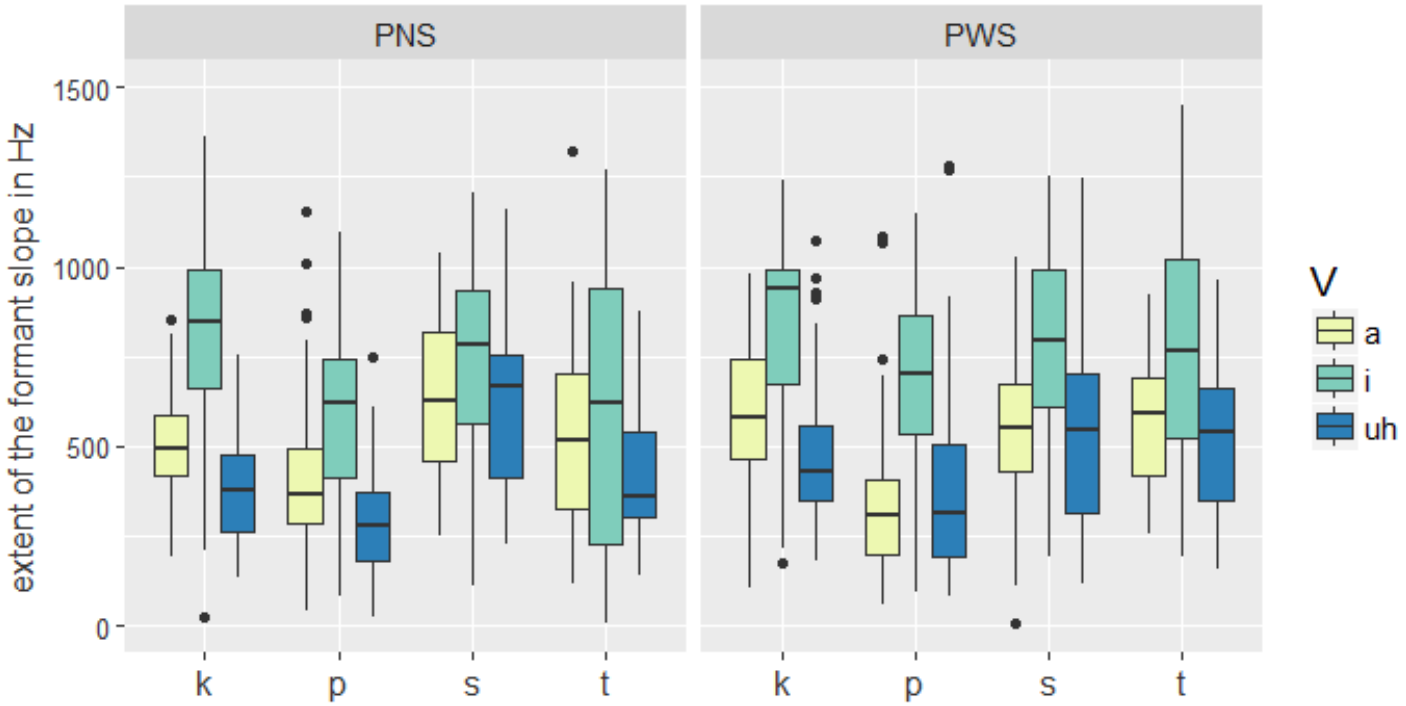


Figure 26 *Formant slope extent for PWS and PNS across consonant (/k, p, s, t/) and vowel (/a, i, ə /) environments*

The null model allowed the intercept and slopes to vary by vowel and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was vowel (i.e., /ɑ, i, ə /). Including vowel improved model fit ($X^2_{(2)} = 8.36$, $p < 0.05$). The reference level was set at /ɑ/. Formant slope extent for /i/ was on average larger than slope extent for /ɑ/ (β = 186.06, SE(β) = 63.87, t = 2.91) while slope extent for /ə/ did not differ statistically from that for /ɑ/ (β = -32.44, SE(β) = 25.01, t = -1.3).

Subsequently we included consonant (/k/, /p/, /s/, /t/) as a fixed effect. The addition of consonant as a fixed effect improved the model fit ($X^2_{(3)} = 126.72$, $p < 0.001$). The reference level was set at /k/. Slope extent for /p/ was on average significantly smaller than that for /k/ (β = -90.86, SE(β) = 8.68, t = -10.47) as was the slope extent for /t/ compared to that for /k/ (β = -35.62, SE(β) = 8.66, t = -4.11). The slope extent for /s/ in contrast did not differ when compared to /k/ (β = -13.72, SE(β) = 8.89, t = -1.54 – see Figure 26).

We next included the interaction between vowel and consonant, which further improved model fit ($X^2_{(6)} = 57.68$, $p < 0.001$). Examination of the model indicated that the slope extent difference between /ɑ/ and /i/ is diminished in /s/ (β = -62.4, SE(β) = 21.43, t = -2.91) and increased in /t/ context (β = 71.07, SE(β) = 21.81, t = 3.37) while there is no effect on that /ɑ/-/i/ difference in /p/ (β = 31.83, SE(β) = 21.37, t = 1.49). The difference in slope extent between /ɑ/ and /ə/ is increased in /p/ (β = 43.44, SE(β) = 20.99, t = 2.07) with no effect in /s/ (β = 12.07, SE(β) = 21.81, t = 0.55) or /t/ context (β = 15.49, SE(β) = 21.14, t = 0.73). Adding speaker group as a fixed effect improved model fit ($X^2_{(1)} = 5.73$, $p = 0.02$). Slope extent was significantly larger for PWS when compared to PNS (β = 69.93, SE(β) = 27.58, t = 2.54).

Adding the interaction for speaker group further improved model fit ($X^2_{(11)} = 112.31$, $p < 0.001$). PWS produced significantly smaller slope extents in /p/ (β = -154.83, SE(β) = 29.74, t = -5.21) and /s/ (β = -146.13, SE(β) = 30.18, t = -4.84) environments. For full details of this model please see Table 24.

Table 24 *Model coefficients (in Hz) for formant slope extent*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	340.99	38.34	8.894	Speaker Intercept	104.88
Vowel (reference level = /ɑ/)				- /i/ vs. /ɑ/	230.48
- /i/ v /ɑ/	140.315	95.568	1.468	- /ə/ vs. /ɑ/	75.16
- /ə/ v /ɑ/	-63.924	42.838	-1.492	Residual	208.24
Consonant (reference level = /k/)					
- /p/ v /k/	-34.593	21.682	-1.595		
- /s/ v /k/	87.483	22.679	3.857		
- /t/ v /k/	-75.221	20.894	-3.600		
Group (reference level = PNS)	90.469	53.690	1.685		
Interaction Vowel : Consonant					
- /i/ : /p/	17.930	31.261	0.574		
- /ə/ : /p/	8.973	30.451	0.295		
- /i/ : /s/	-133.628	31.626	-4.225		
- /ə/ : /s/	2.547	31.827	0.080		
- /i/ : /t/	99.794	29.757	3.354		
- /ə/ : /t/	28.33	17.76	1.595		
Interaction Vowel : Group					
- /i/ : PWS	83.029	134.503	0.617		
- /ə/ : PWS	42.499	59.891	0.710		
Interaction Consonant : Group					
- /p/ : PWS	-154.828	29.735	-5.207		
- /s/ : PWS	-146.134	30.180	-4.842		
- /t/ : PWS	39.145	29.584	1.323		

3.2.3.3 Transition Rate

The measure of formant transition rate was obtained by relating the slope extent to the slope duration (see Figure 27). Longer durations with steady slope extent result in lower transition rates whereas shorter durations with constant extent result in higher transition rates. Equally can transition rates be affected by changes in slope extent where smaller slope extents lead to higher transition rates and larger slope extents to lower slope transition rates. Using linear mixed-effects modelling we investigated the formant transition rates by consonant and vowel environment as well as by speaker group (see Figure 28).

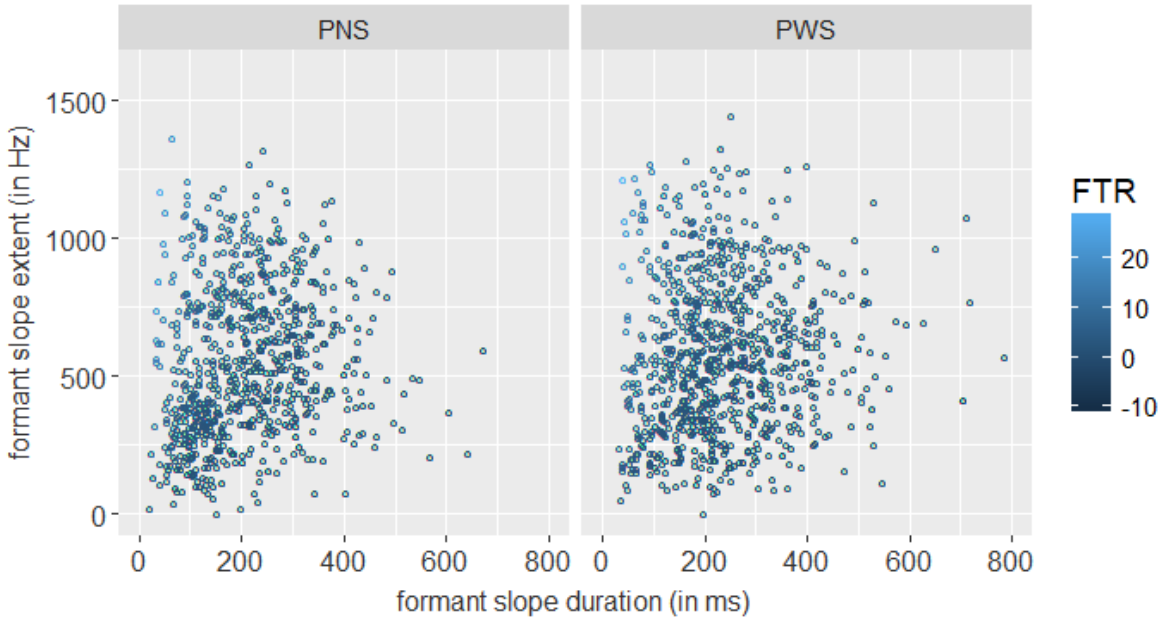


Figure 27 *Slope durations (x-axis) by slope extent (y-axis) indicating the transition rate (colour coded ranging from dark blue to light blue) by speaker group*

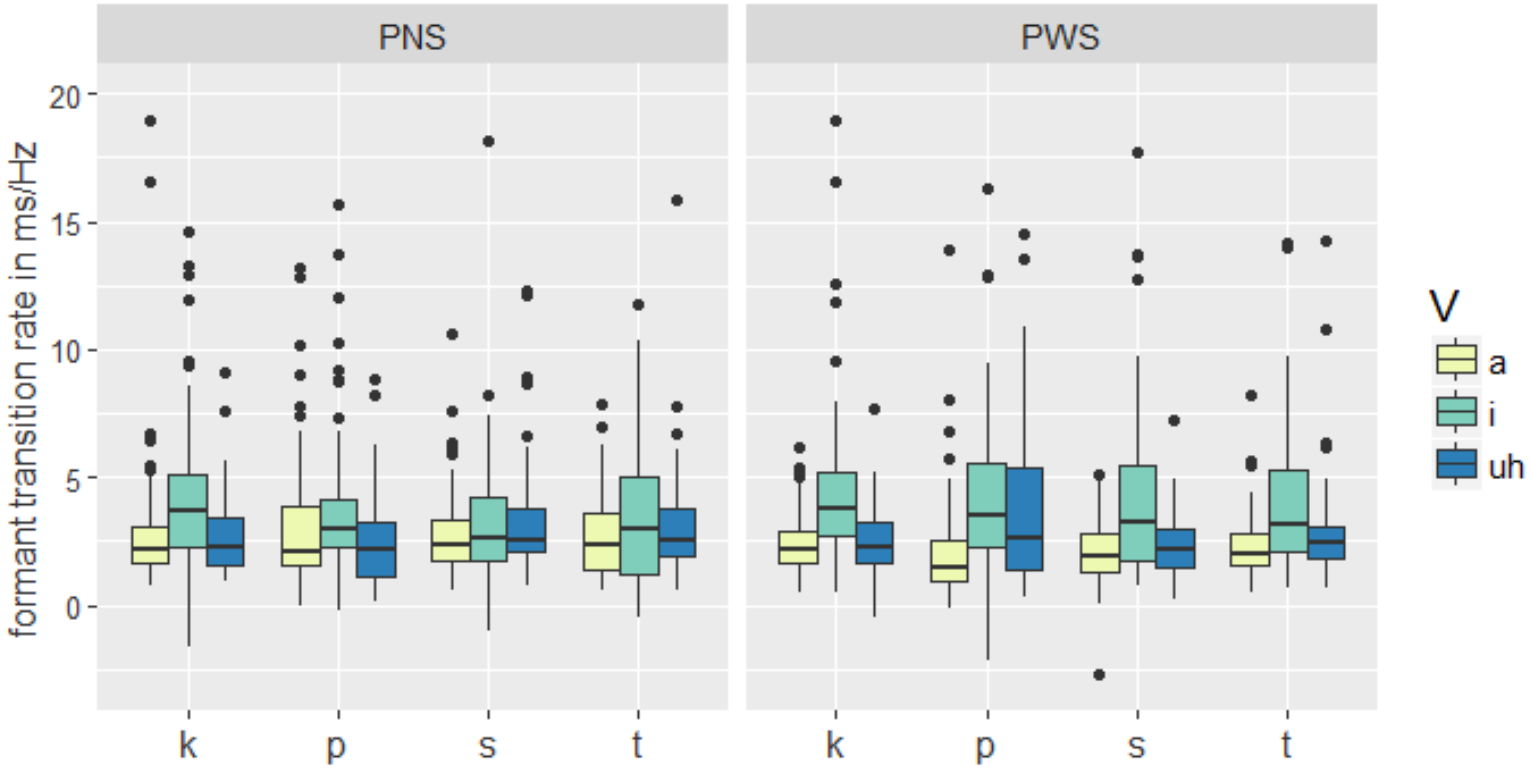


Figure 28 *Formant transition rate for PWS and PNS across consonant (/k, p, s, t/) and vowel (/a, i, ə /) environments*

Table 25 *Model coefficients (in Hz/ms) for formant transition rate*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	1.675	0.147	11.361	Speaker Intercept	0.553
Vowel (reference level = /ɑ/)				- /i/ vs. /ɑ/	1.345
- /i/ v /ɑ/	0.748	0.333	2.244	- /ə/ vs. /ɑ/	0.739
- /ə/ v /ɑ/	0.062	0.200	0.311	Residual	1.5590
Consonant (reference level = /k/)					
- /p/ v /k/	0.113	0.096	1.178		
- /s/ v /k/	-0.460	0.097	-4.764		
- /t/ v /k/	-0.520	0.095	-5.455		
Interaction Vowel : Consonant					
- /i/ : /p/	0.028	0.137	0.204		
- /ə/ : /p/	0.260	0.135	1.932		
- /i/ : /s/	-0.504	0.137	-3.666		
- /ə/ : /s/	0.037	0.140	0.266		
- /i/ : /t/	1.071	0.135	7.927		
- /ə/ : /t/	0.312	0.136	2.304		

The null model allowed the intercept and slopes to vary by vowel and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was vowel (i.e., /ɑ, i, ə /). Including vowel improved model fit ($X^2_{(2)} = 9.09$, $p < 0.01$). The reference level was set at /ɑ/. Formant transition rate for the F2 slope for /i/ was on

average significantly higher than the formant transition rate for /ɑ/ (β = 0.92, SE(β) = 0.3, t = 3.06) while the formant transition rate for /ə/ was comparable to that of /ɑ/ (β = 0.24, SE(β) = 0.18, t = 1.32).

Subsequently we included consonant (/k/, /p/, /s/, /t/) as a fixed effect. The addition of consonant as a fixed effect improved the model fit ($X^2_{(3)}$ = 230.23, p < 0.001). The reference level was set at /k/. Formant transition rate for /p/ was on average higher compared to that for /k/ (β = 0.21, SE(β) = 0.06, t = 3.67) while the formant transition rate for /s/ (β = -0.63, SE(β) = 0.06, t = -10.9) was significant smaller with no difference for /t/ (β = -0.05, SE(β) = 0.06, t = -0.87 – see Figure 28).

We next included the interaction between vowel and consonant, which further improved model fit ($X^2_{(6)}$ = 172.65, p < 0.001). Examination of the model indicated that the formant transition rate difference between /ɑ/ and /i/ is decreased in /s/ context (β = -0.50, SE(β) = 0.14, t = -3.67) but increase in /t/ context (β = 1.07, SE(β) = 0.14, t = 7.93) when compared to /k/ context. In contrast, the formant transition rate difference between /ɑ/ and /ə/ is increased in only /t/ context (β = 0.31, SE(β) = 0.14, t = 2.30) when compared to /k/ context. Adding speaker group as a fixed effect did not improve model fit ($X^2_{(1)}$ = 0.51, t = 0.48). For full details of the model please see Table 25.

3.2.3.4 Variation & Homogeneity

While no differences were found between groups for the measures of slope duration, slope extent or transition rate, speaker groups differed with regards to homogeneity for both measures of slope duration and slope extent for prompts where C=/k, p, s, t/ (see Figure 29).

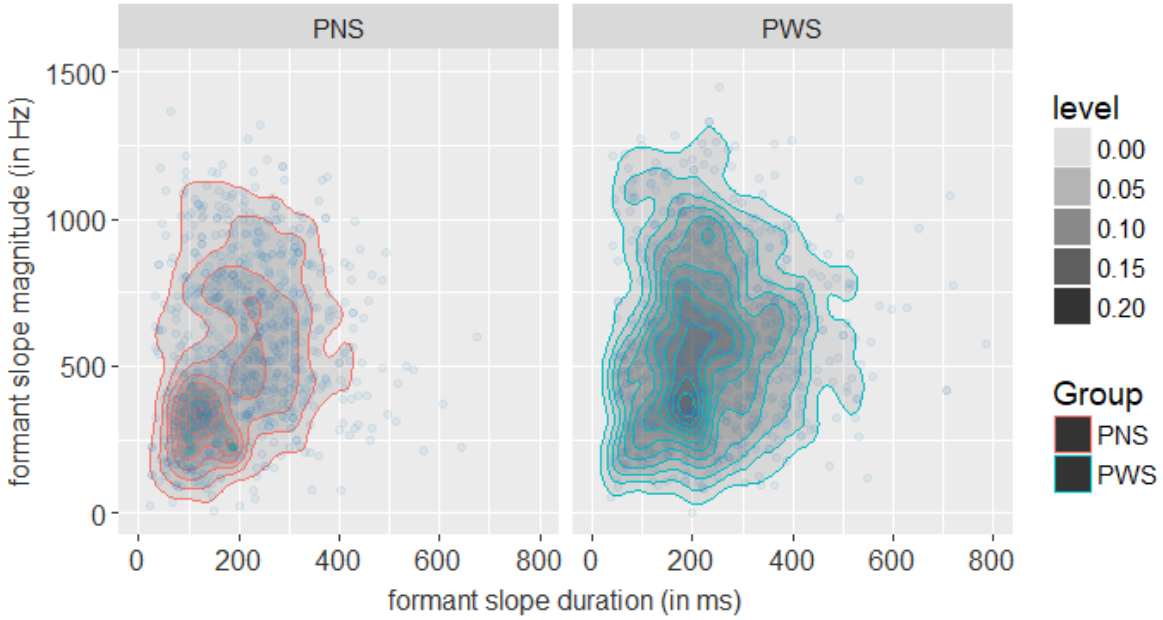


Figure 29 *Density plot for slope durations (x-axis) by slope extent (y-axis) indicating the transition rate (colour coded ranging from dark blue to light blue) by speaker group*

We explored the homogeneity of variance of the two speaker groups regarding the measures of F2 slopes presented above. We applied the Fligner-Killeen Test of Homogeneity of Variance to prompts where C = /k, p, t, s/. We ran the test separately for slope duration and slope extent (as dependent variable) with speaker group as independent variable. The Fligner-Killeen Test of Homogeneity of Variance returned significant differences for the variation of the two speaker groups for both slope durations ($X^2_{(1)} = 3.87$, $p < 0.05$) as well as slope extent ($X^2_{(1)} = 17.12$, $p < 0.001$) and transition rate ($X^2_{(1)} = 15.49$, $p < 0.001$; see Table 26). This significant difference allows us to reject the null hypothesis that F2 slope durations, slope extent or the transition rates are homogenous for the two speaker groups.

In a second step, we applied the coefficient of variation to the measures of slope duration, the measures of slope extent and the formant transition rate – each by speaker group (see Table 26). The coefficient of variation returned comparable

variances for the two speaker groups on slope duration (PNS: 45.72%; PWS: 42.86%). The coefficient of variation on slope extent in contrast returned a meaningfully larger variation for PNS (86.44%) when compared to that of PWS (67.32%; see Table 26) resulting overall in a more variation in transition rate for PNS (103.01%) when compared to that of PWS (93.33%).

Table 26 *Coefficient of variation and Homogeneity of Variance for F2 slope durations, slope extent and slope transition rate by speaker group.*

	homogeneity of variance	coefficient of variation (in %)	
	PWS vs. PNS	PNS	PWS
slope duration	$p < 0.05$	45.72	42.86
k	$p = 0.39$	45.27	37.80
t	$p < 0.001$	38.58	39.06
s	$p = 0.12$	31.08	32.81
p	$p = 0.23$	52.62	49.50
slope extent	$p < 0.001$	86.44	67.32
k	$p = 0.15$	77.02	59.10
t	$p < 0.001$	123.27	72.00
s	$p < 0.001$	63.14	55.77
p	$p < 0.001$	66.58	78.40
slope FTR	$p < 0.001$	103.01	93.33

3.2.4 Summary

Three acoustic measures were applied to investigate consonant-vowel transitions in the fluent speech of people who stammer and their control speakers. First, we employed linear mixed effects models to explore mean durations of acoustic segments (i.e., consonantal closure, release and vowel) to see whether at this level differences between speaker groups could already be observed. Next, we compared locus equations for the two speaker groups where the steepness of the regression line is indicative of the degree of coarticulation. Third, we obtained measures of

formant slope and compared the slope duration, extent and transition rate for the two groups.

3.2.4.1 Group Means

Comparing the two speaker groups, segment mean durations reveal a significant difference where PWS perform with consistently longer mean durations compared to PNS. This effect can be observed independent from consonant or vowel environment. The longer durations for PWS suggest an overall lower speed when compared to PNS.

To see whether the lower speed is echoed by a more global measure, we examined the full utterance length as well as the speech onset time, i.e., the delay from the speech-triggering beep to the onset of speech. For statistical investigation, we used linear mixed effects models. With group as fixed effect, we found a significant difference where PWS produced overall longer utterance durations ($\beta = 79.00$, $SE(\beta) = 39.49$, $t = 2.00$) when compared to typical speakers. This was expected as the individual segments were also longer in PWS speech.

Regarding the time it took speakers to initiate speech, we found a delayed speech onset for PWS ($\beta = 0.09$, $SE(\beta) = 0.04$, $t = 2.11$). This delay, however, should be considered with caution as speakers were not instructed to produce the utterances as quickly as possible. We therefore refrain from drawing inferences with regard to typical reaction times obtained in reaction time studies.

Using locus equations or formant slopes, differences between speaker groups were not as easily identified. Regarding locus equations, the two speaker groups differed regarding the intercept where PWS were shown to have lower intercept overall indicating a lower degree of coarticulation when compared to PNS. Slope steepness, in contrast, did not reveal any differences between groups. Further, no differences in locus equation is found for the different degrees of severity of a stammer. The latter may be due to the sample size not having enough power to return statistical significance.

Formant slopes revealed a difference in slope extent between the two speaker groups with larger slope extent for PWS when compared to PNS. For slope duration and transition rate speaker group was regarded as not having explanatory value.

3.2.4.2 Group Variance

In addition to mean measures we also investigated variability of measures as an indicator of stability of motor programs. For acoustic segment durations we observed differences for the two speaker groups for both coefficient of variance (in %) and homogeneity of variance. Homogeneity of variance returns a significant difference for both speaker groups. Overall, PWS perform with increased variance, which appears to be driven mainly by differences in mean durations for closure segments where, proportionally, the difference is largest.

Similar to the group differences observed in the variance of segment duration, locus equation and formants slope variation are also indicative of two speaker groups. For locus equations, we explored the variation for onset and target values for both groups. The coefficient of variation returned less variation for PWS when compared to PNS. The statistical test of homogeneity of variance further returned overall significant group differences in homogeneity for both F2 onset and F2 target values thereby confirming that these values are not homogeneous and must stem from two separate data pools. The overall significant difference appears to be driven mainly by differing variation in lingual consonant environments while no group difference was found for labial consonant environments.

The results of the Fligner-Killeen test further returned a significant difference for the duration and extent of the F2 slopes – again reflecting the two speaker groups. Proportionally, the group of PWS shows less variation in both formant slope measures, namely, slope duration and slope extent – consequently also lowering the variance in formant transition rate for PWS.

3.2.4.3 CV Segments

Utterances were designed to systematically cover a range of places and manners of articulation, which would inevitably affect CV transition patterns. Results indicate that acoustic mean durations differ as a function of the vowel where durations are significantly shorter for /ə/ environment compared to /ɑ/ environment across all consonants, i.e., /k, p, s, t/. The shorter segment durations can be explained by the shorter distance that needs to be covered when targeting a central vowel as compared to the longer distance associated with the corner vowels /ɑ/ and /i/. Moreover, schwa as a central vowel is likely to coarticulate, reducing articulatory effort for schwa in contrast to that of corner vowels (Browman & Goldstein, 1992b).

Locus equations as anticipated differed as a function of consonant. Results show the highest intercept and steepest slope for bilabial /p/, followed by /k/ and alveolar /s/ and /t/ with the highest intercept and flattest slope. This pattern was consistent for both speaker groups. These differences for consonants were shown to be statistically significant.

For formant slope durations and extent, as well as the resulting transition rate overall, a significant difference was found for vowel and for consonant. As would be expected, findings support a direct relation between slope duration and extent. An increased slope duration is reflected in larger slope extent, which we observed for /s/, while shorter durations correspond to decreased slope extent as observed for bilabial /p/. Transition rate was highest for /p/ and lowest for /s/ with /k/ and /t/ ranging between them.

3.3 Preliminary Discussion of Acoustic Findings

Revisiting the questions posed at the outset of this thesis (see section 1.6), results show that even the perceptually fluent utterances from PWS and PNS contain characteristics that distinguish the two speaker groups as is implied by Wingate's Fault-Line hypothesis (Wingate, 1988; see section 1.4.2.1). Perceptually, the

productions of PWS and PNS appear to be fluent and are indistinguishable. Results, however, show that these perceptually indistinguishable productions have acoustic characteristics that allow to distinguish the fluent productions of PWS from those of PNS. These differences show that acoustic analysis has the capability to uncover events that lie beyond perception.

PWS present with overall longer and more variable segment durations when compared to PNS – a difference that could be shown for each consonant separately. The longer durations suggest slower speech rate, which was echoed with the global measures of utterance duration and measure of durations for speech initiation (see 3.2.4.1).

Increased speech rate is typically associated with increased motor demand while lower speed is associated with decreased demand on motor control. In PWS lower speech rate is subsequently associated with enhanced fluency (Andrews et al., 1982). This is supported by findings from Andrade and colleagues who have shown an inverse correlation between stammer severity and speech rate (Andrade, Cervone, & Sassi, 2003). Their findings show decreased speech rates with increasing stammer severity. Slower speech rate in PWS was also previously observed (Borden et al., 1987; Guitar, 2013; Postma, Kolk, & Povel, 1990).

Lower speech rate further is associated with both longer durations and increased variability. Smith and Sugarman (B. L. Smith, Sugarman, & Long, 1983) investigated the causality between segment duration and variability in CWS. Their findings indicate that the increased variability is both (a) a statistical consequence of the longer durations as well as (b) an indicator of motor control deficit.

The slowed speech production could therefore be a strategy which PWS employ to remain stable in their productions (Onslow et al., 1992) and to maintain fluency. Slower execution has often been claimed to provide more control over execution (Nudelman, Herbrich, Hoyt, & Rosenfield, 1987; Tasko, Mcclean, & Runyan, 2007). It could, however, also be reflective of the speaker's increased sensitivity in

monitoring (1.4.1.3) especially since speakers were aware of their speech being investigated at the time of recording (Postma, 2000).

In addition to the slower speed, the two speaker groups presented with a significant difference in formant slope extent. PWS performed with greater slope extents, which indicates less coarticulation when compared to PNS. Coarticulation merges adjacent segments and brings together their formant frequencies. A lack of coarticulation, in contrast, increases the formant distance that needs to be covered in transition. The greater slope extent therefore suggests that PWS coarticulate less. Adjacent segments maintain their formant quality resulting in increased formant differences that need to be covered in a transition. This is also in line with the longer segment durations we observed for PWS.

Locus equations were sensitive to differences in coarticulation as a function of the consonant. The pattern broadly follows that proposed by Recasens (Recasens, 1985) where tongue body involvement increases coarticulatory resistance. Labial consonants like /p/ would therefore be expected to show largest degree of coarticulation, which is represented with the steepest locus equation slope as could be shown. Differences between groups showed flatter slopes for PWS suggesting less coarticulation when compared to PNS – a finding in line with the results obtained using formant slopes.

Looking at measures of variance, segment durations for the two speaker groups differ significantly. PWS perform with overall increased variance, which would be expected in the context of overall longer durations. Unexpected however was the finding of larger variation on closure durations. The increased variance in closure durations may be indicative of remainders of disfluencies in the speech of PWS. While these disfluencies were not perceptually salient during the fluency judgement (section 2.5.2) they can be captured using acoustic measures.

In contrast to the increased variance for durations, results show decreased variance for locus equations as well as for formant slope measures. For locus equations, PWS

perform with overall decreased variance on both onset and target measures. Parallel to the overall smaller variance for locus equations we observe less variance in PWS for all three slope measures (duration, extent and formant transition). The decrease in variance may suggest that PWS have generally narrower targets leaving less flexibility when transitioning between adjacent consonant and vowel segments, which supports Wingate's prediction where the tight coupling of CV gestures contributes to the complexity of the transition (see 1.4.2.4).

For locus equations, the decreased variance was observed for only lingual consonant environments, which will be relevant to the following section where we employ lingual kinematics to explore the fluent speech of PWS and PNS in CV transitions.

4 Articulatory Analysis

In this chapter we present the articulatory analysis, including methodology, results and a brief discussion of those results. The methodology section (section 4.1) presents information on data treatment, including splining of the ultrasound frames (section 4.1.1), the fixation of a measurement vector (section 4.1.2) and landmarking (section 4.1.3). Subsequently, we will present the measures applied (section 4.1.3.33.1.3), which comprise measures of stroke duration, displacement and peak velocity. Results (section 4.2) for these measures and their implications will be discussed (section 4.3) before we turn to the general discussion and conclusion (chapter 5).

We present the articulatory analysis of the fluent CV recordings where C = /k/ followed by /ɑ, i, ə/. All recordings were produced in isolation in the typically voiced condition. Analysis was performed on all recordings categorised as fluent (see section 2.5.2 for more information). Analysis was limited due to time constraints and constraints imposed by the nature of ultrasound imaging with most of the relevant tongue surface being visible in the production of velar sounds and less so in the production of bilabial or alveolar sounds.

We recorded ultrasound data (Wrench, 2015) to explore articulatory events in the fluent speech of PWS. The ultrasound frames relative to the acoustic signal were investigated. Following the acoustic landmarking (see 3.1.1), we splined the ultrasound frames relative to the acoustic signal of the V_PCV utterance, i.e., from the beginning of the prothetic schwa to the end of the vowel in the CV sequence.

Articulatory segmentation was based on the articulatory trajectory of the tongue surface. The trajectory included the tongue movement from a stable position into consonantal closure and following the closure into the release towards the vowel where the tongue reaches a relatively stable position and the cycle starts again (Boucher, 2008). Subsequently, we will refer to the lingual movement towards lingo-palatal closure as (gestural) onset, whereas the movement trajectory away

from this closure will be referred to as (gestural) offset. Frames were splined, and displacement and velocity were measured along the 'best fitted' fanline. Measures for onset and offset were obtained and analysed across speaker group and vowel context. Additionally, session effects were controlled for to preclude any effects of training.

4.1 Methodology

4.1.1 Splines

We splined the ultrasound frames to extract the tongue surface location from the ultrasound image (Iskarous, 2005a; Li, Kambhamettu, & Stone, 2005; Unser & Stone, 1992). This mechanism extracts the relevant data for analysis while other structures also detected by the ultrasound scanner are removed for the following data analysis (see the online tutorial for annotations and splines (Articulate Instruments Ltd, 2015)). Each ultrasound frame consists of greyscale pixels that construct a fan shaped image. The colour of the pixels represents the strength of reflection of differences in the density of tissue. These differences in density help to identify the tongue surface and other structures by the ultrasound signal. Depending on the strength of the signal, a thick diffuse white line indicates the tongue surface.

To encode information about the tongue contour location, ultrasound frames need to be spline-fitted and splined. The software cannot distinguish white traces resulting from the tongue surface from those resulting from other structures. It is therefore important to manually mark / correct the relevant traces of the tongue surface, which is referred to as splining. The researcher inserts reference points for the tongue surface onto the ultrasound image. These points are connected by the software resulting in a line that approximates the tongue surface – thereby excluding any structures other than the tongue. The connected tongue surface points are then read by the software and used for analysis.

For the current project, splines were inserted using a semi-automatic edge-detection function in AAA (Articulate Instruments Ltd, 2012). Splines were inserted and semi-automatically aligned using an in-built function in the software. Alignment is based on a fan-shaped measurement grid, with 42 control points ("knots"). Each knot sits on a radial fanline, and its location can be set manually or using the semi-automatic edge-tracker. A knot is defined by its fan-number and the distance from the origin of the fan.

4.1.1.1 Spline Templates

Tongue movements are measured against the reference of the passive articulators. Using stabilisation, information about the oral cavity size and passive articulators are assumed to be consistent across frames for each speaker. To ensure that the ultrasound remained stable over the duration of the recording session, splining templates are created containing the splines of the palate for each speaker and recording session (see section 2.3.2.2). When creating a template, the researcher first defines the rough shape of the palate looking at the recording of the participant swallowing water (Shaker, 2009). Swallowing is a process where the tongue inevitably moves along the palate. While sliding along the palate, the white / near-white pixels that represent the tongue surface indirectly map a trace of the palate. The palate shape indicates the ceiling for the tongue movement. Templates can help to verify the stability of the passive articulators within and across recording sessions.

In this study, the palate trace was used to ensure that the ultrasound probe was fixated and that data within the same session could be compared against another. Data were recorded in two sessions with a longer break during which speakers took off the headset. This means that the headset was put on again for the second recording session. In most cases the probe was not located at the same place as in the first recording session. Using the palate trace from the templates, data were aligned so they could be collapsed for analysis.

In addition to the palate spline, two splines are inserted to loosely indicate the roof and the floor of the mouth. These roof and floor splines are used as a search window for the semi-automatic splining mechanism. The search window allows to exclude structures other than the tongue. The tongue surface is fit next.

The tongue is the only active articulator where the spline needs to be adjusted in each frame that is spline-fitted (see section 4.1.1.3). To capture the full articulatory movement of the tongue from beginning to end of the V_PCV utterance, frames were splined beginning with the first frame in the schwa preceding the consonant and ending with the last frame of the second vowel of the target stimulus. Starting at the first keyframe in the schwa, the template was loaded carrying splines for the search window (roof and floor of the mouth), the palate and the tongue surface. For all subsequent frames only the tongue was adjusted, which was done tracking the edge (i.e., the best dark-to-light edge in the ultrasound image) of the tongue contour semi-automatically.

4.1.1.2 Edge Tracking

Splines are mathematical functions that are used for fitting curves. Edge Tracking is a mechanism that allows for semi-automated tracking on the frames displaying the relevant tongue movement (i.e., the lingual movement from the schwa sound into the consonant and transitioning from the consonant onto the vowel). Each frame is based on a fan shaped grid with 42 radial measurement and control lines. The first tongue spline is manually fitted to the shape of the tongue by defining a given number of points along the length of the tongue. After manually fitting, it is refined semi-automatically using an inbuilt “snap to fit” function, a local search which scans locally to the input along each of the 42 fan lines for the best dark-to-light edge.

Spline fitting was performed with a confidence multiplier of 45. The confidence multiplier indicates the validity of a signal in a series of up to 42 knots. In cases where confidence multiplier was set to the maximum of 100, the software attempts to find a tongue trace even if no signal from the tongue is available. In cases, where the confidence multiplier is set at too low of a value, existing signal is not

recognised and would not be available for later analysis. It is therefore important to adjust the confidence multiplier to be most sensitive to the signal from the tongue while not being overly sensitive and capturing artefacts. It is possible to manually correct the edge by moving incorrect knots. When correcting, it is particularly useful to set confidence to 0% at the anterior and posterior edges of the tongue surface in the image to prevent any false positives.

The fitted first spline is the starting point for an automated edge tracking of the tongue surface contour throughout the subsequent frames of the recording. The tracking function bases a new spline in a following frame on the shape finalised for the preceding frame and does a snap-to-fit local search for the new edge. This process continues and cycles through all subsequent frames always fitting the spline based on the previous one. Local search edge tracking works well for tracking dynamic changes in the mid-sagittal curve where each frame is slightly different to the one preceding it. Tracking the splines throughout the frames is based on a combination of edge detection and brightness detection. Tracking was interrupted and new starting point for the splines was defined manually if artefacts in the ultrasound image led the automated tracking to go astray.

4.1.1.3 Spline-Fitting

Spline-fitting refers to the process where the number of ultrasound frames is determined that will inform the later analysis. In cases where the articulatory analysis is driven by the acoustic signal, a single frame relative to the acoustic event (e.g., burst) may suffice. In cases, however, where larger amounts of frames need to be investigated to find the region of interest, splines are inserted into frames for the duration of the ultrasound sequence that is of interest. In this project, the consonant-vowel transition in V_PCV sequences is of interest. We inserted splines relative to the acoustic signal beginning with the first frame of the first vowel through to the last frame of the second vowel.

The density / rate of frames that one decides to spline for a sequence can be crucial. While too many frames take much longer to spline and to analyse, too few

frames may skip over details that could be useful to address certain articulatory questions. In this study the ultrasound data was recorded at a high framerate of 121 frames per second. The analysis software AAA allows to spline fractions of the originally recorded data. We decided to spline a fraction of the recorded 121 frames per second, which reduced the time for data processing and analysis considerably.

4.1.1.3.1 Processing Time vs. Quality

Several factors played a role in the decision by how much the data could be reduced in terms of the framerate splined and analysed. The factors included the processing time (the time it takes to spline and track the splines) and the quality of the analysis, but also the technical limitations set by the Edge Tracking (Frisch, 2010; Iskarous, 2005a; Li et al., 2005). Reducing the rate of splined frames to every second or third frame reduces the processing time by approximately the same fraction. A reduction in processing time was indispensable given the time limitations for the current project. However, reducing the number of splined frames also affects the quality of the analysis in that it limits the detail of information that can be obtained from the data. This inevitably has implications for the results.

4.1.1.3.2 Search Window

A more apprehensible factor perhaps was the limitations set by the edge tracking mechanism. After having manually fitted and refined the tongue contour in the first frame of a recording, the tracker takes a copy of the tongue contour spline into the next frame to be splined. Based on that copy the tracker searches within a predefined search window (the default setting here is 10% of the circumference of the ultrasound image) for the best dark-to-light edge where it places the spline for the frame. The newly created spline is then copied into the next frame where it serves as a reference for the edge tracker (Wrench, 2015) searching for the most pronounced dark-to-light edge. This process is repeated over and over for every keyframe until the end of the tracked sequence of keyframes.

The search area for fitting a spline was crucial as it is the area within which the tracker searches for the most distinct dark-to-light edge. While the search area needed to be kept narrow enough to ensure precise fitting of the tongue contour, it needed to be wide enough to capture the movement of the tongue contour from one frame to another. With larger tongue movement, the search window needed to be wider to capture the movement. The wider the search window, however, the larger the chances that the edge tracker detects artefacts which may distract the tracker. With too many distractions, the tracker loses its ability to fit splines properly to the tongue surface.

4.1.1.3.3 Frame Rate for Splining

With a reduced number of frames, the movement between them increases. And the reduction in frames behaves disproportionally to the amount of movement, meaning, the more frames we dropped, the larger the movement from one frame to another. This made it more difficult for the tracker, sometimes even impossible, if the amount of tongue movement surpasses the limitations set by the search area which is approximately 10% of the circumference of the ultrasound image. Exceeding the search window causes the tracker to capture artefacts instead of the tongue surface. When tracking through multiple frames, the edge tracker then takes a copy of the spline that has gone astray and take this as a reference when looking for the tongue contour in the next frame. This corrupts the automated tracking process rendering it invalid to analysis.

4.1.1.3.4 Summary

The number of frames to be splined was an essential consideration in balancing the requirements for the tracker and the amount of time it would take for the tracking process. We decided to leave the size of the search window at the default setting of 10% and to manipulate the number of frames.

After weighing all options and several trials of automatically tracking the tongue contour we decided to reduce the amount of data by splining every third frame of

the defined interval. The displacement of the tongue surface between four frames has proven to be too large to be captured by the edge tracker (Iskarous, 2005a). The tongue surface position moves too much that it exceeds the search area of the edge tracker that is applied when tracking and splining multiple frames. Every time the edge tracker loses track because artefacts are picked up instead of the tongue surface, manual corrections are required. The amount of manual corrections is inversely proportional to efficiency. The larger the number of manual corrections, the lower the efficiency. The compromise between processing speed and processing quality consisted in tracking every third frame, which resulted in a more manageable amount of data that could be processed faster while also maintaining decent processing quality due to the trackability. The decrease in the number of splined frames meant an increase in the number of manual corrections. Splines that led the tracker to not 'find' following splines were corrected.

4.1.2 Measurement Vector

Once the tongue splines were fitted, kinematic information of the tongue surface could be extracted from the splines. The movement could be measured on any of the 42 fanlines where the tongue surface was visibly crossing. For each speaker and recording session, a fanline / polar coordinate was chosen along which the tongue moving towards the palate was measured (Heyne & Derrick, 2015). Out of the 42 fanlines, the fanline in the velar area was chosen as the measurement vector along which the extent of tongue surface displacement was largest (see section 5.2.1.2).

The splines corresponding to the acoustic signal of the CV utterance were superimposed (see Figure 30 panel a). The superimposed splines create a 2D image informing about the extent of tongue surface displacement along radial distances originating at the ultrasound probe (see Figure 30 panel b). The radial distances covered a field of view (FOV) of 135° of the midsagittal plane informing about crossing points of the tongue surface at different points in the vocal tract, indirectly reflecting differential movement displacement of different areas on the tongue

surface (Iskarous, 2004). The resulting image was used to estimate the strength of the signal as well as to establish areas on the tongue surface where displacement is largest and where the measurement vector would be placed (see Figure 30 right panel).

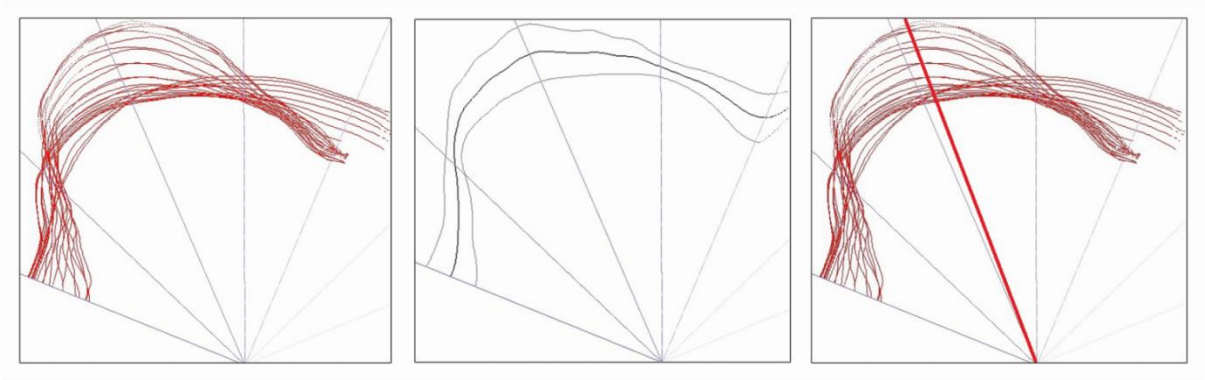


Figure 30 *Tongue splines with measurement vector along the fanline with greatest relative displacement (Heyde, Scobbie, et al., 2016)*

For all measurements presented here, the candidate vectors were all fan radii. The vector chosen was based on objective criteria. A mean spline was created based on the values for each spline at each of the 42 fan radii, with standard deviation and confidence indicated. All subsequent measurements were taken along the scan line with the largest value of standard deviation from the mean spline. A relatively high confidence of the AAA spline fitting (at least 85% overall) was used as a threshold for the validity of data.

4.1.3 Articulatory Landmarking

Relative to the acoustic signal of the CV utterance, measures of displacement and absolute velocity were extracted. Both displacement and velocity were calculated using the motion of the tongue surface spline along the chosen measurement vector. To verify that the correct measurement vector was chosen, the displacement of the tongue surface along neighbouring fanlines was investigated (see Figure 31). The raw data was smoothed using a smoothing function within the

AAA software (Wrench, 2015) to obtain bell-shaped trajectories (S. G. Adams, Weismer, & Kent, 1993) for the tongue approaching and moving away from the palate. Because data smoothing is highly significant and may affect findings, potential limitations are discussed in section 5.2.1.3.

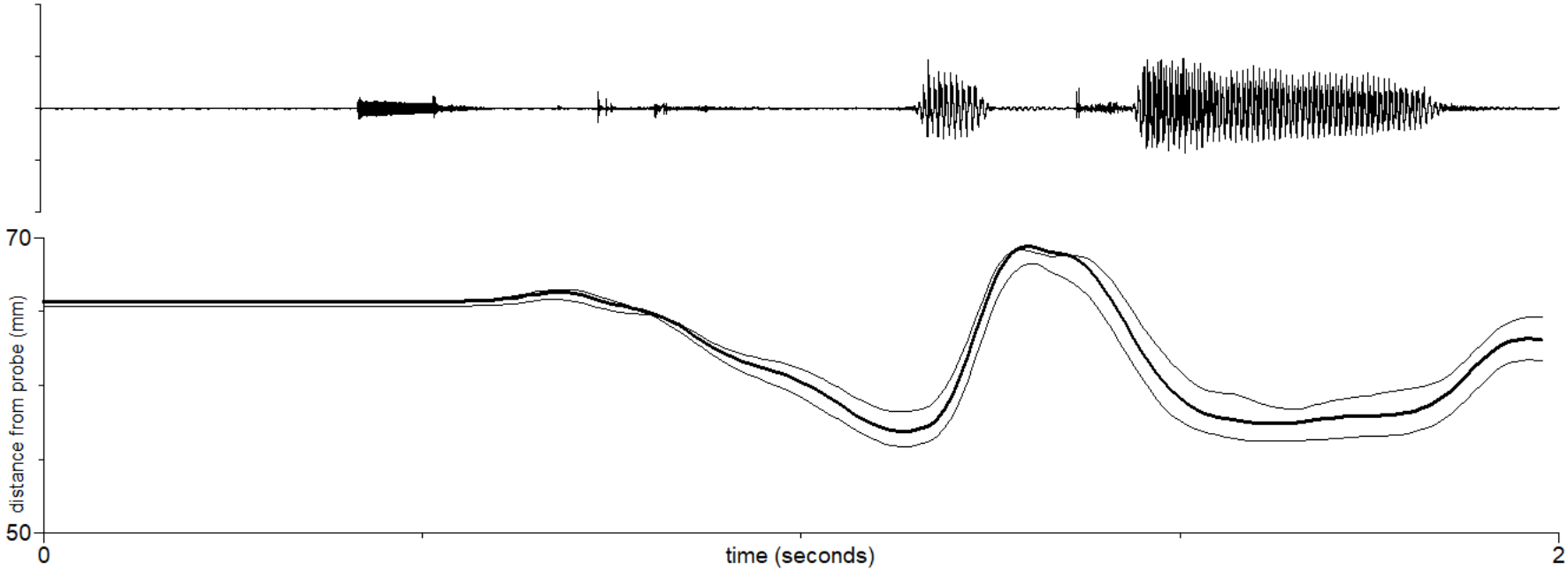


Figure 31 *Acoustic signal (upper panel) and displacement of the tongue surface along the measurement vector placed at three neighbouring fanlines (lower panel) for the production of /*ə kɑ/.

4.1.3.1 Onset and Offset

Within the CV utterance particularly two articulatory movements were of interest to our study, which we will refer to as onset and offset strokes in the following. The term ‘stroke’ refers back to Tasko and Westbury’s definition of “a period between two successive local minima in the speed history of an articulatory point” (Tasko & Westbury, 2002, p 127). The first of the two articulatory movements is the articulatory onset or closing phase. During this phase, we observe the movement of the tongue from the neutral starting point at schwa towards the palate for consonantal closure. The onset movement starts off at the schwa with very low velocity which increases until a maximum is reached. Following the maximum, velocity decreases until the tongue reaches the palate where minimum velocity is reached before a new cycle (with velocity increases again) is initiated for the second articulatory movement.

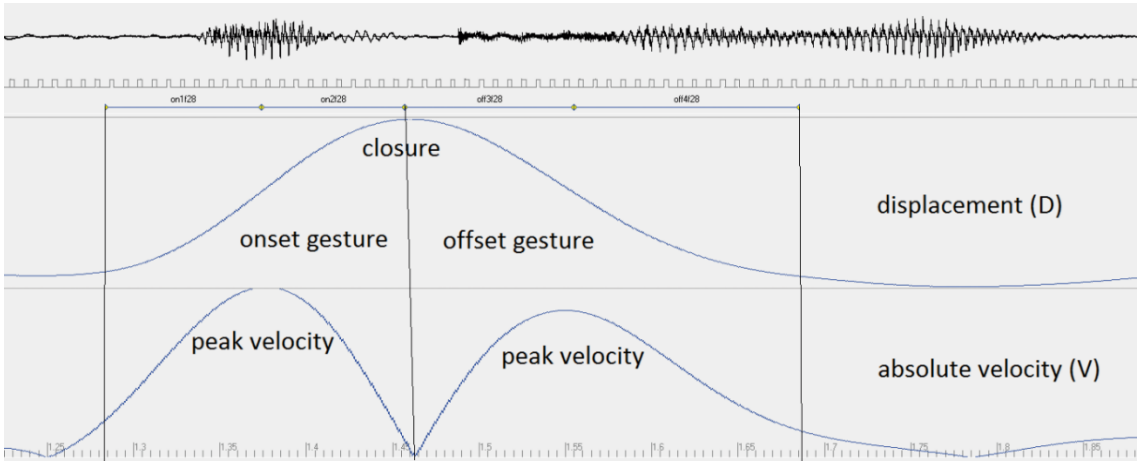


Figure 32 *Displacement and velocity curves for the tongue surface movement along the measurement vector*

The second articulatory movement is the offset or release phase, which describes the movement of the tongue away from the palate towards a stable position in the subsequent vowel. Starting at the palate velocity increases, reaches its maximum and decreases before reaching a stable position within the vowel (see Figure 32 for details).

4.1.3.2 Threshold Measure

We employed a 20% threshold in the beginning of onsets and at the end of offsets to determine boundaries consistently (rationale and implications discussion in section 5.2.1.4). Beginning and end of the articulatory movement strokes are marked with a low absolute velocity. The two movement strokes of onset and offset are directly adjacent, and their boundary is marked by a clear minimum in the absolute velocity trace when the tongue surface touches the palate. The beginning of the onset and the end of the offset stroke are also marked with a low absolute velocity both when starting off and ending in a stable articulatory position. These boundaries, however, may be less clear with velocity traces trailing off to both ends.

We adopted a threshold measure that is traditionally applied in articulatory research (Kroos, Hoole, Kühnert, & Tillmann, 1997; Tasko & McClean, 2004 for additional information on threshold criterion). The threshold measure takes the maximum velocity of a movement stroke and takes 20% of that value to determine where the movement stroke begins (onset) or ends (offset). This way, signal and noise are distinguished, and onset and offset durations can be determined systematically. Applying the threshold measure, we embrace the fact that lingual kinematic data may not always be smooth and categorical but may contain irregularities. The measure simplifies highly complex articulatory movements and allows for consistent segmentation.

4.1.3.3 Smoothing

Once the data of the tongue displacement moving along the measurement vector is extracted, it needs to be smoothed to remove artefacts. This is a process known from postprocessing of EMA data (1.5.3.2). At this stage it is crucial to determine the right smoothing factor. Raw data captures all levels of movement of the tongue – not only the directed movement of the tongue body along the vector. Even tiny pulses are captured in raw data. With all kinds of directed and non-directed movements, the resulting data becomes quite distorted / fuzzy / wiggly.

For the current study, the signal was smoothed using the 4th order polynomial Savitsky-Golay function at 200ms width in the Math analysis values tool integrated in AAA (Articulate Instruments Ltd, 2012). Smooting was required to to cancel out noise which makes it possible to visualise and extract the directed movement patterns (see section 5.2.1.3).

4.1.4 Articulatory Measures

4.1.4.1 Stroke Duration

Duration measures were extracted for the two movement strokes (onset and offset) where the tongue surface moves perpendicular along the measurement vector. Onset strokes refer to movements of the tongue surface away from the ultrasound probe while offset strokes refer to movements towards the ultrasound probe. Because we placed the vector in the velar area to investigate movements towards and away from velar closure, onsets are equivalent to trajectories approaching the palate for velar closure with offsets representing movement trajectories in the release phase away from the palate.

Velocity minima and the 20% velocity threshold (Fuchs et al., 2006; Kühnert et al., 2006; Mooshammer et al., 2012; Pouplier & Waltl, 2008; Shaw & Hoole, 2011) were used to mark onset and offset boundaries which we used to extract durations for the movement strokes. Boundary information were extracted automatically using the export function in the ultrasound software AAA (Wrench, 2015). These were then read into R (RStudio Team, 2015) to obtain durations relative to onset and offset trajectories.

4.1.4.2 Displacement

The movement of the tongue surface (i.e., the spline) relative to the measurement vector was further translated into measures of motion. The first measure is that of displacement which indicates the radial distance from the origin at the ultrasound probe to where the spline crosses the vector. Within AAA (Wrench, 2015) it is possible to visualise and extract these measures of displacement where movement

towards the palate (onset) results in positive values while movements away from the palate (offset) results in negative values. Calculating the absolute value for these displacement values was then noted as the distance of the tongue surface movement along the vector.

4.1.4.3 Peak Velocity

The displacement values are further translated into values of absolute velocity, which indicate how quickly (mm/s) the tongue surface moves along a particular fanline. Movements into the consonantal constriction (onset) and movements away from consonantal constriction into the subsequent vowel (offset) were distinguished. Both movement strokes were defined with low velocity on either end. The low velocity in the beginning and at the end of a movement stroke indicates the stable position of the articulator, i.e., the tongue, before initiating movement and when reaching the target position. Between these two points of initiating movement and reaching the target position, the velocity increases and reaches a maximum. This is the case for both movement strokes.

From the velocity measures were obtained maximum velocity values for onset and offset trajectories separately. Measures for maximum velocity were extracted at the turning point between increasing and decreasing velocity for both the movement of the tongue surface towards (onset) and away from (offset) palatal contact.

4.1.4.4 Average Speed

In addition to the peak velocity measure we investigated average speed of the tongue surface moving along the vector. To obtain the average speed, the extent of a movement stroke was extracted from the absolute values of tongue displacement and related to the duration for that movement stroke. The resulting values indicate the average speed of the tongue surface travelling along the vector to reach the respective target.

4.2 Results

Using ultrasound, onset and offset strokes were measured as the tongue surface moved along the measurement vector. Both movement strokes were measured regarding their duration and the maximum velocity achieved.

We analysed data from 9 PWS and 9 control speakers after excluding three experimental speakers (speakers 3, 9, 11) and their control speakers (speakers F and G) in addition to one speaker who was excluded from this study due to the strong influence of his L1 (speaker 4).

We present the analysis for 644 V_PCV utterances with 637 fluent and 7 disfluent recordings with a focus on the fluent productions. Utterances combine the prothetic schwa in the first vowel position (V_P = /ə/) with a voiceless velar closure as consonant (C = /k/) followed by three vowel conditions in the vowel position (V = /ɑ, i, ə /). Utterances are balanced with respect to the vowel (V) and speaker group (PWS vs. PNS). The materials include 222 recordings where V = /ɑ/, 197 recordings with V = /i/ and 225 recordings where V = /ə/. Half of the material was produced by people who stammer (308 recordings) and the other half (336 recordings) by their control speakers.

4.2.1 Stroke Duration

For the analysis of the articulatory data all recordings that were judged disfluent were treated separately from those judged as being perceptually fluent in the MFC perception study (Boersma & Weenink, 2015). Duration measures were obtained for the two movement strokes for the tongue surface moving towards (i.e., onset) and away from the palate (i.e., offset). Measures were obtained for the two speaker groups also differentiating by vowel.

4.2.1.1 Descriptive Analysis

Pooling across speaker group and vowel we learn that offsets in the fluent articulatory data are typically longer and more variable (M = 176.77ms; SD = 57.70ms) compared to onset movements (M = 140.07ms; SD = 29.22ms; see Table 27). The same holds when looking at the two speaker groups separately both presenting with significantly shorter onset compared to offset durations (PWS x onset: M = 146.11ms; SD = 29.71ms, PWS x offset: M = 194.05ms; SD = 61.74ms, PNS x onset: M = 134.65ms; SD = 27.71ms; PNS x offset: M = 161.49ms; SD = 49.13ms).

Table 27 *Duration measures (in ms) for movement direction (onset / offset) by vowel (/ɑ, i, ə/) and speaker group (PNS / PWS)*

		overall		/ɑ/		/ə/		/i/	
		Onset	Offset	Onset	Offset	Onset	Offset	Onset	Offset
All fluent	Mean	140.07	176.77	141.33	196.92	142.02	191.76	136.48	136.62
	SD	(29.22)	(57.70)	(27.88)	(50.94)	(25.35)	(54.04)	(34.14)	(48.22)
PWS fluent	Mean	146.11	194.05	143.70	210.77	145.70	211.01	149.64	149.36
	(SD)	(29.71)	(61.74)	(27.54)	(52.63)	(24.80)	(56.28)	(37.08)	(58.00)
PNS fluent	Mean	134.65	161.49	139.06	183.57	138.50	173.36	126.28	127.33
	(SD)	(27.71)	(49.13)	(28.13)	(45.66)	(25.47)	(44.84)	(27.82)	(37.21)
PWS disfl		197.93	230.93						
		(72.20)	(117.40)						

A comparable relationship with longer and more variable offsets (M = 230.93ms; SD = 117.40ms) compared to shorter and less variable onset durations (M = 197.97ms; SD = 72.20ms) can be observed for the disfluent productions from PWS. Comparing fluent and disfluent recordings both mean durations and standard deviations (for onsets as well as offsets) are clearly larger in the disfluent recordings (N = 7).

4.2.1.2 Statistical Analysis

Articulatory data were analysed using linear mixed effects models. Data were modelled in milliseconds. All models included the maximal justified random effects structure. We took a forward stepwise approach when adding fixed effects predictors in order to allow us to explore effects from a theoretical basis. At each step model fit was compared to that of the previous model in order to determine whether the additional predictor improved model fit (i.e., had explanatory value). Following the statistical analysis using linear mixed effects models, we explore variance and homogeneity for the two speaker groups. We first report articulatory durations followed by articulatory peak velocity measures.

Table 28 *Model coefficients (in ms) for articulatory onset and offset stroke duration*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	106.752	3.758	28.403	Speaker Intercept	493.72
Stroke type				- /ki/ vs. /kɑ/	283.66
(reference = onset)	26.92	2.909	9.254		
Group				- /kə/ vs. /kɑ/	201.88
(reference = PNS)	14.396	5.259	2.737		
Interaction				Session Intercept	1261.75
offset: PWS	20.666	4.240	4.874		
				- /ki/ vs. /kɑ/	798.19
				- /kə/ vs. /kɑ/	1.23
				Residual	1419.41

In a first step, we compared durations for onset and offset stokes in a single model. The null model allowed the intercept to vary by session with varying slopes for session by vowel as well as by speaker with slopes varying for speaker by vowel (i.e.,

as random effects). Including stroke type (onset, offset) as a fixed effect improved model fit ($X^2_{(1)}$ = 261.61, p < 0.001). As predicted, offset strokes were significantly longer when compared to onset strokes (β = 36.62, SE(β) = 2.14, t = 17.13). Subsequently we added group as a fixed effect, which improved model fit ($X^2_{(1)}$ = 20.32, p < 0.001) revealing that on average PWS produce longer movement strokes than PNS (β = 24.63, SE(β) = 4.83, t = 5.1). Finally, we added the interaction of stroke type (onset, offset) and speaker group (PWS, PNS), which further improved model fit ($X^2_{(1)}$ = 23.53, p < 0.001). The interaction shows that PWS produce longer offset strokes when compared to PNS (β = 20.67, SE(β) = 4.24, t = 4.87). For full details of this model, please see Table 28.

Table 29 *Model coefficients (in ms) for articulatory onset stroke duration*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	141.232	3.206	44.05	Speaker Intercept	399.3
				- /ki/ vs. /kɑ/	405.3
				- /kə/ vs. /kɑ/	5.717
				Session Intercept	0.0
				- /ki/ vs. /kɑ/	1.776
				- /kə/ vs. /kɑ/	3.426
				Residual	392.7

Next, we treated onset and offset trajectories in two separate models to test the hypotheses that predicts that speaker groups do not differ in onset strokes, while PWS and PNS differ in offset movement strokes.

The null model allowed the intercept to vary by session with varying slopes for session by vowel as well as by speaker with slopes varying for speaker by vowel (i.e., as random effects). Including speaker group (PNS, PWS) as a fixed effect did not improve model fit ($X^2_{(1)}$ = 2.49, p = 0.11). As predicted, the two groups did not

receive explanatory value to describe onset stroke durations. For full details of this model please see Table 29.

Table 30 *Model coefficients (in ms) for articulatory offset stroke duration*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	181.105	12.242	14.794	Speaker Intercept	1397.55
Vowel				- /ki/ vs. /kɑ/	1739.58
(reference = /kɑ/)					
- /ki/ v /kɑ/	-54.726	9.516	-5.751	- /kə/ vs. /kɑ/	1091.51
- /kə/ v /kɑ/	-4.026	6.975	-0.577	Session Intercept	169.46
				- /ki/ vs. /kɑ/	54.21
Group				- /kə/ vs. /kɑ/	13.38
(reference = PNS)	28.424	9.085	3.129		
				Residual	1239.33

Subsequently, we tested the statistical difference between the two speaker groups for offset stroke durations. Parallel to the onset duration testing, the null model allowed the intercept to vary by session with varying slopes for session by vowel as well as by speaker with slopes varying for speaker by vowel (i.e., as random effects).

The first predictor to be included as a fixed effect was vowel (i.e., /ɑ, i, ə /). Including vowel improved model fit ($X^2_{(2)} = 6.96$, $p < 0.05$). As would be anticipated, offset movement durations for /i/ were on average shorter than those for /ɑ/ ($\beta = -55.71$, $SE(\beta) = 9.673$, $t = -5.759$) with no significant difference for stroke durations between /ɑ/ and /ə/ ($\beta = -4.01$, $SE(\beta) = 6.939$, $t = -0.578$). Subsequently we included the main variable of interest, speaker group (PWS, PNS) as a fixed effect.

The inclusion of speaker group as fixed effect further improved model fit ($X^2_{(1)} = 8.935$, $p < 0.01$). Inspection of the model indicated that on average durations were

significantly longer for PWS compared to PNS (β = 28.424, SE(β) = 9.085, t = 3.129). Finally, we added the interaction between vowel and speaker group. The inclusion of that interaction did not improve model fit ($X^2_{(2)}$ = 0.793, p = 0.673). For full details of this model please see Table 30.

4.2.1.3 Variance & Homogeneity

We applied the Flinger-Killeen test, to compare homogeneity of variance between speaker groups (PWS, PNS) for stroke types (onset, offset). In a second step, we employed coefficients of variation to compare the relative variation for onset and offset stroke durations by speaker group and vowel context (see Table 31).

Table 31 Variation and Homogeneity for onset and offset stroke durations for both speaker groups across the different vowel environments

	onset			offset		
	homogeneity of variance	coefficient of variation		homogeneity of variance	coefficient of variation	
	PNS vs. PWS	PNS	PWS	PNS vs. PWS	PNS	PWS
overall	p = 0.30	20.33	20.58	p < 0.001	30.42	31.81
/ɑ/	p = 0.11	20.89	20.45	p = 0.06	33.91	39.08
/i/	p = 0.59	20.67	27.54	p < 0.001	27.63	43.07
/ə/	p = 0.53	18.91	18.42	p < 0.01	33.30	41.80

Regarding the measure of homogeneity (Fligner-Killeen Test of Homogeneity of Variance) we found no evidence for differences in variance in onset stroke durations between speaker groups ($X^2_{(1)}$ = 1.06, p = 0.03). This was also reflected when exploring variances in homogeneity by vowel where we found no significant difference between speaker groups (/ɑ/: $X^2_{(1)}$ = 2.49, p < 0.11; /i/: $X^2_{(1)}$ = 0.29, p = 0.59; /ə/: $X^2_{(1)}$ = 0.40, p = 0.53). The non-significant difference in homogeneity of variance is also reflected in the coefficient of variation which is similar for the two speaker groups indicating similar degrees of variation – even for individual vowel

environments. Offset stroke durations in contrast behave differently for the two speaker groups. The Fligner-Killeen Test of Homogeneity of Variance indicates differences for the two speaker groups overall ($X^2_{(1)} = 16.95$, $p < 0.001$). The overall significant difference in homogeneity of variance is also found for two out of three vowel environments (/i/: $X^2_{(1)} = 16.14$, $p < 0.001$; /ə/: $X^2_{(1)} = 7.27$, $p < 0.01$). The third vowel environment shows a trend towards significance (/ɑ/: $X^2_{(1)} = 2.94$, $p = 0.06$).

Employing the coefficient of variation, we further see that the degree of variation is larger for offset stroke durations (32.64%) when compared to onset stroke durations (20.86%). In contrast to onset stroke durations, offset stroke durations show a slightly higher relative variation for PWS (31.81%) when compared to PNS (30.43%). This effect can be observed across vowel environments where the coefficients of variation are consistently higher for the people who stammer when compared to typical speakers (/ɑ/: PNS = 33.91%, PWS = 39.08%. /i/: PNS = 27.63%, PWS = 43.07%, /ə/: PNS = 33.30%, PWS = 41.80%).

4.2.2 Peak Velocity

Analogous to the analysis of the durations, we analysed 637 fluent and 7 disfluent recordings from 9 PWS and 9 control speakers after excluding seven speakers due to L1 influence strongly affecting the English pronunciation, poor ultrasound imaging quality or redundant control speakers.

Looking at the speakers' peak velocity measures overall it becomes apparent that higher mean values are achieved in onset movement strokes (M = 117.70 mm/s) when compared to offsets (M = 84.20 mm/s). This contrast of higher peak velocity in onset movements when compared to offset movements holds for both groups, though it is more prominent in people who stammer. The same is true for the recordings that were categorised as disfluent (onset: M = 110.11 mm/s; offset: M = 98.18 mm/s; see Table 32).

4.2.2.1 Descriptive Analysis

Table 32 *Peak velocity measures (in mm/s) for onset and offset strokes by vowel context and speaker group*

		overall		/ɑ/		/ə/		/i/	
		Onset	Offset	Onset	Offset	Onset	Offset	Onset	Offset
All fluent	Mean (SD)	117.70 (35.62)	84.20 (44.87)	117.39 (33.86)	98.47 (44.47)	212.01 (35.64)	88.71 (44.54)	114.33 (37.31)	62.72 (37.24)
PWS fluent	Mean (SD)	115.19 (38.68)	70.65 (40.10)	115.49 (37.05)	83.79 (37.63)	116.18 (39.08)	73.26 (40.39)	113.56 (40.52)	49.81 (34.50)
PNS	Mean (SD)	119.95 (32.52)	96.13 (45.51)	119.20 (30.57)	112.50 (46.10)	125.63 (31.49)	103.48 (43.45)	114.93 (34.79)	72.15 (36.48)
PWS disfl		110.11 (52.47)	98.18 (28.40)						

What can be noted regarding the standard deviation of onset and offset trajectory is that (with only the exception of the disfluent recordings) the trajectory away from consonantal closure (offset) is usually less stable with a larger standard deviation (SD = 44.87) compared to that of the onset trajectory (SD = 35.62).

Comparing peak velocity measures across the two groups of PWS and PNS we observe a clear group difference where PNS reach higher peak velocities for both trajectories towards and away from consonantal closure (onset: M = 119.95 mm/s, offset: M = 96.13 mm/s) compared to those of PWS (onset: M = 115.19 mm/s, offset: M = 70.65 mm/s).

The difference between onset and offset mean peak velocity varies with group. While the peak velocity difference between onset and offset in PWS reaches approximately 44.54 mm/s, the difference is considerably lower in PNS with only 23.82 mm/s. The importance of these group differences is supported by statistical testing (see Table 33).

4.2.2.2 Statistical Analysis

As with the analysis of stroke durations, we took a forward stepwise approach when adding fixed effects predictors in order to allow us to explore effects from a theoretical basis. At each step model fit was compared to that of the previous model in order to determine whether the additional predictor improved model fit (i.e., had explanatory value).

Table 33 *Model coefficients (in ms) for peak velocity in onset and offset*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	104.86	3.629	28.89	Speaker Intercept	932.778
Stroke type				- /i/ vs. /ɑ/	564.271
(reference = onset)	-33.827	1.616	-20.93		
				- /ə/ vs. /ɑ/	64.659
				Session Intercept	389.207
				- /i/ vs. /ɑ/	427.552
				- /ə/ vs. /ɑ/	9.715
				Residual	826.929

First, we compared peak velocity for onset and offset stokes in a single model. The null model allowed the intercept to vary by session with varying slopes for session by vowel as well as by speaker with slopes varying for speaker by vowel (i.e., as random effects). Including stroke type (onset, offset) as a fixed effect improved model fit ($X^2_{(1)} = 371.89$, $p < 0.001$). As predicted, the peak velocity for offset strokes was significantly lower when compared to that of onset strokes ($\beta = -33.83$, $SE(\beta) = 1.62$, $t = -20.93$). Subsequently we added group as a fixed effect, which did not improve model fit ($X^2_{(1)} = 3.23$, $p = 0.07$) indicating that overall the two speaker groups do not differ in peak velocity.

Next, we treated onset and offset trajectories in two different models to test the hypothesis that predicts that a) speaker groups do not differ in onset strokes, while b) PWS and PNS differ in offset movement strokes which constitute the transition from consonantal closure to the subsequent vowel.

For peak velocity in onset trajectories, the null model allowed the intercept to vary by session with varying slopes for session by vowel as well as by speaker with slopes varying for speaker by vowel (i.e., as random effects). Including speaker group (PNS, PWS) as a fixed effect did not improve model fit ($X^2_{(1)} = 0.27$, $p = 0.6$). As predicted, no significant difference for the two groups was found for the peak velocity of onset movement strokes. For full details of this model please see Table 34.

Table 34 *Model coefficients (in ms) for peak velocity in onset*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	141.971	4.729	24.31	Speaker Intercept	653.10
				- /i/ vs. /ɑ/	143.88
				- /ə/ vs. /ɑ/	80.37
				Session Intercept	13.29
				- /i/ vs. /ɑ/	18.81
				- /ə/ vs. /ɑ/	12.87
				Residual	518.41

Next, we tested the statistical difference between the two speaker groups for peak velocity in offset stroke movements. Parallel to the peak velocity testing in onsets, the null model allowed the intercept to vary by session with varying slopes for session by vowel as well as by speaker with slopes varying for speaker by vowel (i.e., as random effects).

Table 35 *Model coefficients (in ms) for peak velocity in offset*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	111.608	7.815	14.282	Speaker Intercept	1420.26
Vowel (reference = /ɑ/)				- /i/ vs. /ɑ/	2439.92
- /i/ v /ɑ/	-38.861	8.737	-4.448	- /ə/ vs. /ɑ/	289.29
- /ə/ v /ɑ/	-10.255	3.405	-3.012	Session Intercept	8.45
				- /i/ vs. /ɑ/	2.97
Group (reference = PNS)	-25.077	7.912	-3.170	- /ə/ vs. /ɑ/	0.03
				Residual	385.39

The first predictor to be included as a fixed effect was vowel (i.e., /ɑ, i, ə /). Including vowel improved model fit ($X^2_{(2)} = 6.67$, $p < 0.04$). Peak velocity values in offset movements for /i/ ($\beta = -37.54$, $SE(\beta) = 8.87$, $t = -4.23$) as well as for /ə/ ($\beta = -10.27$, $SE(\beta) = 3.40$, $t = -3.02$) were on average lower than those for /ɑ/. Subsequently we included the main variable of interest, speaker group (PWS, PNS) as a fixed effect.

The inclusion of speaker group as fixed effect further improved model fit ($X^2_{(1)} = 9.47$, $p < 0.01$). Inspection of the model indicated that on average peak velocity values were significantly lower for PWS compared to PNS ($\beta = -25.08$, $SE(\beta) = 7.91$, $t = -3.17$). Finally, we added the interaction between vowel and speaker group. The inclusion of that interaction did not improve model fit ($X^2_{(2)} = 0.31$, $p = 0.86$). For full details of this model please see Table 35.

4.2.2.3 Variance & Homogeneity

We applied the Flinger-Killeen test, to compare homogeneity of variance in peak velocity between speaker groups (PWS, PNS) for stroke types (onset, offset). In a second step, we employed coefficients of variation to compare the relative variation for onset and offset stroke peak velocity by speaker group and by vowel context.

Table 36 *Variation and Homogeneity for onset and offset stroke peak velocities for both speaker groups across the different vowel environments*

	onset			offset		
	homogeneity of variance	coefficient of variation		homogeneity of variance	coefficient of variation	
	PNS vs. PWS	PNS	PWS	PNS vs. PWS	PNS	PWS
overall	$p < 0.01$	27.11	33.58	$p = 0.09$	47.34	56.75
/ɑ/	$p = 0.12$	25.48	30.89	$p = 0.05$	38.43	31.37
/i/	$p = 0.18$	29.00	33.78	$p = 0.89$	30.41	28.76
/ə/	$p = 0.02$	26.25	32.58	$p = 0.14$	36.22	33.67

The measure of homogeneity (Fligner-Killeen Test of Homogeneity of Variance) provides evidence for group differences for the variance of peak velocity reached in onset strokes ($X^2_{(1)} = 8.24$, $p < 0.01$). Homogeneity of variance in peak velocity did however not differ for individual vowel environments (/ɑ/: $X^2_{(1)} = 2.48$, $p < 0.12$; /i/: $X^2_{(1)} = 1.76$, $p = 0.18$; /ə/: $X^2_{(1)} = 5.60$, $p = 0.02$). The coefficients of variation are relatively similar with a tendency for lower variance in the peak velocity for PNS. In offset movement strokes, homogeneity of variances does not suggest a difference between groups with regards to peak velocity variance. In contrast to onset peak velocity variances, variances in offset peak velocity are consistently lower for PWS. For details, please see Table 36.

4.2.3 Distance

Distance measures were extracted from the articulatory movement data capturing the tongue surface moving along the vector. Movement distance from the stable position at schwa until the palate is captured in onset values, while movement distance of the tongue surface moving towards the ultrasound probe along the vector is described in offset values (see Table 37).

4.2.3.1 Descriptive Analysis

Table 37 *Distance (in mm) for 'onset' and 'offset' strokes by vowel context and speaker*

		overall		/ɑ/		/ə/		/i/	
		Onset	Offset	Onset	Offset	Onset	Offset	Onset	Offset
All	Mean	9.44	8.80	9.49	10.72	9.88	9.66	8.86	5.41
	(SD)	(3.27)	(4.77)	(3.12)	(3.90)	(2.77)	(4.53)	(3.88)	(4.25)
PWS	Mean	9.47	8.39	9.66	10.11	9.65	8.87	9.02	5.39
(fluent)	(SD)	(3.27)	(4.71)	(3.24)	(3.57)	(2.90)	(4.55)	(3.73)	(4.98)
PNS	Mean	9.41	9.16	9.33	11.28	10.12	10.47	8.72	5.42
	(SD)	(3.29)	(4.81)	(3.02)	(4.12)	(2.65)	(4.41)	(4.03)	(3.68)
PWS		10.94	12.65						
(disfl.)		(4.92)	(3.02)						

Pooling the data for speaker group and vowel we learn that the distance covered along the vector is typically larger and less variable in onset (M = 9.44 mm; SD = 3.27 mm) compared to offset movements (M = 8.80 mm; SD = 4.77 mm). The same holds when looking at the two speaker groups separately both presenting with larger distance at lower variability in onset (PWS x onset: M = 9.47 mm; SD = 3.27 mm; PNS x onset: M = 9.41 mm; SD = 3.29 mm) compared to offset movements (PWS x offset: M = 8.39 mm; SD = 4.71 mm; PNS x offset: M = 9.16 mm; SD = 4.81 mm)

An inverse relationship with lower distance at larger variability in onsets (M = 10.94 mm; SD = 4.92) compared to larger distance at lower variability in offset movements (M = 12.65; SD = 3.02) can be observed for the disfluent productions from PWS. Comparing the fluent and disfluent recordings the mean distance values (for onsets as well as offsets) are visibly larger in the disfluent recordings compared to the fluent recordings, while standard deviations are comparable for fluent and disfluent recordings.

4.2.3.2 Statistical Analysis

Table 38 *Model coefficients (in mm) for articulatory movement distance*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	10.921	7.488	15.965	Speaker Intercept	5.091
Variable (reference = onset)	-1.277	3.644	-1.838	- /ki/ vs. /kɑ/	2.960
Vowel (reference = /kɑ/)				- /kə/ vs. /kɑ/	0.886
- /ki/ v /kɑ/	-5.500	6.659	-0.713	Session Intercept	0.059
- /kə/ v /kɑ/	-1.277	4.245	1.518	- /ki/ vs. /kɑ/	0.003
Interaction Variable : Vowel				- /kə/ vs. /kɑ/	0.020
- onset : /ki/	4.700	0.585	8.028	Residual	8.623
- onset : /kə/	1.410	0.551	2.557		

The null model allowed the intercept and slopes to vary by session and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was movement direction (i.e., onset / offset). Including movement direction improved

model fit ($X^2_{(1)}$ = 6.273, p < 0.05). On average distance was lower for offsets when compared to onsets (β = 0.585, SE(β) = 0.246, t = 2.378).

Subsequently we included vowel (/ɑ, ə, i/) as a fixed effect. The addition of vowel as a fixed effect improved the model fit ($X^2_{(2)}$ = 8.486, p < 0.05). The reference level was set at/ɑ/. Distance for /i/ was on average significantly lower than that for /ɑ/ (β = -3.100, SE(β) = 0.4323, t = -7.172) while there was no significant difference between /ɑ/ and /ə/ (β = -0.558, SE(β) = 0.336, t = -1.661).

We next included the interaction between movement direction and vowel, which further improved model fit ($X^2_{(2)}$ = 63.088, p < 0.001). Examination of the model indicated that the difference in distance between onset and offset was decreased for /i/ as compared to /ɑ/ (β = -4.700, SE(β) = 0.585, t = 8.028) as well as for /ə/ compared to /ɑ/ (β = 1.410, SE(β) = 0.551, t = 2.557).

We now further included the main variable of interest, speaker group (PNS, PWS) as fixed effect. The inclusion of speaker group as fixed effect did not improve model fit ($X^2_{(1)}$ = 0.257, p = 0.612). For full details of this model please see Table 38.

4.2.4 Average Speed

Relating the movement distance to the duration the tongue surface takes to cover that distance, we obtained a measure of average speed for the onset and offset movement strokes.

4.2.4.1 Descriptive Analysis

Pooling the data for speaker group and vowel we learn that the average speed with which the tongue moves along the vector is typically larger and less variable in onset (M = 69.01 mm/s; SD = 21.87 mm/s) compared to offset movements (M = 50.42 mm/s; SD = 23.68 mm/s). The same holds for the mean values of the two speaker groups separately. Both groups present with larger average speed mean values in onset (PWS x onset: M = 67.43 mm/s; PNS x onset: M = 70.12 mm/s) compared to offset movements (PWS x offset: M = 43.87 mm/s; PNS x offset: M =

56.39 mm/s). A comparable relationship with larger average speed values in onsets (M = 65.39 mm/s) compared to lower average speed values in offset movements (M = 58.92 mm/s) can be observed for the disfluent productions from PWS.

Table 39 *Average Speed for 'onset' and 'offset' strokes by vowel context and speaker*

		overall		/ɑ/		/ə/		/i/	
		Onset	Offset	Onset	Offset	Onset	Offset	Onset	Offset
All fluent	Mean (SD)	69.01 (21.87)	50.42 (23.68)	68.69 (20.37)	57.36 (24.19)	70.57 (20.64)	51.67 (23.45)	66.94 (25.10)	40.85 (24.57)
PWS	Mean (SD)	67.43 (23.88)	43.87 (21.07)	69.10 (22.28)	50.19 (19.37)	68.47 (23.11)	42.69 (20.57)	64.04 (26.80)	37.19 (22.11)
PNS	Mean (SD)	70.12 (20.01)	56.39 (26.50)	68.31 (18.55)	63.78 (26.36)	72.63 (17.87)	60.82 (22.80)	69.39 (23.55)	43.54 (26.12)
PWS disfl		65.39 (33.07)	58.92 (22.95)						

Variation in the fluent recordings overall does not differ noticeably. Concerning the two speaker groups, however, an inverse relation can be observed where PWS present with slightly higher average speed values in onset (SD = 23.88 mm/s) when compared to offsets (SD = 21.07 mm/s). In PNS average speed onset values show less variability (SD = 20.01 mm/s) when compared to offsets (SD = 26.50 mm/s)

Comparing the fluent and disfluent mean values, the difference between mean onset and offset average speed values appears to be smaller in the disfluent recordings when compared to the fluent recordings. Further, noticeably larger variation can be observed in the disfluent onset values (SD = 33.07 mm/s) when compared to fluent onset values (SD = 21.87 mm/s). For full details please see Table 39.

4.2.4.2 Statistical Analysis

The null model allowed the intercept and slopes to vary by session and speaker (i.e., as random effects). The first predictor to be included as a fixed effect was movement direction (i.e., onset / offset). Including movement direction improved model fit ($X^2_{(1)} = 145.8$, $p < 0.001$). On average speed was higher for onset when compared to offsets ($\beta = 18.055$, $SE(\beta) = 1.403$, $t = 12.87$).

Subsequently we included vowel (ɑ, ə, i) as a fixed effect. The addition of vowel as a fixed effect did not improve the model fit significantly ($X^2_{(2)} = 5.147$, $p = 0.076$). We now further included the main variable of interest, speaker group (PNS, PWS) as fixed effect. The inclusion of speaker group as fixed effect did not improve model fit ($X^2_{(1)} = 3.405$, $p = 0.065$). For full details of this model please see Table 40.

Table 40 *Model coefficients (in mm/s) for articulatory average speed*

Fixed Effect	**Estimate**	**SE**	**t**	**Random Effect**	**Variance**
Intercept	54.149	2.963	18.28	Speaker Intercept	248.241
Variable (reference = onset)	18.055	1.403	12.87	- /ki/ vs. /kɑ/	231.539
				- /kə/ vs. /kɑ/	1.228
Group (reference = PNS)	-8.691	4.652	-1.868	Session Intercept	0.000
				- /ki/ vs. /kɑ/	108.805
				- /kə/ vs. /kɑ/	13.889
				Residual	313.226

4.2.5 Summary

We applied a novel technique to obtain kinematic measures of transitions in the fluent speech of PWS and PNS. Materials consisted of CV utterances (C=/k/, V= /ɑ, i, ə/). Transitions were broken down into two parts: onset and offset. Onset strokes

describe the lingual movement into consonantal closure and offset strokes describe the movement away from the closure configuration into the subsequent vowel. Along a measurement vector we obtained two acoustic measures, namely, duration and peak velocity for onset and offset movement strokes.

4.2.5.1 Stroke Type

Results show a general positive relation between stroke duration and stroke distance. Both measures are inversely correlated to peak velocity. Statistical analysis returned overall longer offset durations when compared to onset durations, which also reflects the longer distance covered in offset strokes. Onset strokes in contrast are shorter – partially because they are produced at higher peak velocity when compared to offset strokes. This may be accounted for through the ballistic nature of these articulatory movements.

A similar duration-distance relationship was found with regards to vowel differences in offsets where shorter distances are also reflected by shorter durations. As would be expected, measures showed a strong interaction between vowel and offset stroke duration reflecting the different nature of the vowels. When transitioning from /k/ to the high vowel /i/, for example, movement durations were shorter. The shorter durations reflect the shorter distance between their centres when compared to the distance travelled from /k/ closure to the low vowel /ɑ/. No difference in duration could be found for /ɑ/ and /ə/ environments. As was expected, there was no such interaction for onset movement durations and vowel reflecting the sameness of that movement owing to the consistent V_P and C.

In addition to the established positive relationship between movement duration and distance covered, we found an inverse relation between peak velocity and movement distance, as would be expected. Lower duration values are associated with higher peak velocity values and higher duration values with lower peak velocity values. The overall longer offset durations are achieved at lower peak velocity while the shorter onset durations present with higher peak velocity.

4.2.5.2 Group Means

In addition to these expected effects, data further revealed differences between the two speaker groups where PWS performed with overall longer stroke durations when compared to the control group. Results further reveal a significant interaction for speaker group and stroke type where PWS perform with significantly longer offset strokes when compared to PNS.

Looking at onset and offset movement strokes separately, no difference was found for the duration of onset strokes relating to the two speaker groups. In contrast to onset durations, offset durations differed as a function of speaker group where durations were overall longer for PWS when compared to PNS. PNS present with overall shorter durations that are achieved at higher peak velocity values while PWS in contrast present with longer durations at lower peak velocity. This finding can partially be related to the longer acoustic durations found for PWS suggesting overall lower speech rate (see 3.2.4.1), while it does not reflect in a significant difference for the two speaker groups on average speed. As would be expected, distance measures are comparable for both speaker groups.

4.2.5.3 Group Variance

Parallel to the non-significant differences for speaker group in onset duration, no difference was found for variance in onset durations. For offset durations, in contrast, PWS perform at larger variance throughout (see Coefficient of Variance for offset durations; Table 31) also reaching significance.

Concerning variance in peak velocities, the opposite is observed. In onset peak velocity, variances for the two speaker groups vary meaningfully while variances for peak velocity in offset strokes are homogenous.

4.3 Preliminary Discussion of Articulatory Findings

Parallel to the acoustic results (see sections 3.2 and 3.3), the articulatory methodology presented reveals kinematic characteristics that distinguish

perceptually indistinguishable productions from PWS and PNS further supporting Wingate's Fault-Line hypothesis (Wingate, 1988; see section 1.4.2.1).

Adding to the acoustic analyses, the kinematic analysis allowed to distinguish two types of kinematic movement. First, the movement stoke when approaching articulatory closure for /k/ (onset) and second, the movement stroke corresponding to the transition from articulatory closure for /k/ to the subsequent vowel /ɑ, i, ə/ (offset). While for onsets no difference between the two speaker groups was expected, in offsets we expected to find differences for PWS that would confirm the transition deficit hypothesis put forward by Wingate. These expectations were confirmed by the results indicating significant differences for the two speaker groups in neither duration nor peak velocity for onset strokes. For offset strokes, in contrast, results reveal differences for the two speaker groups where PWS produced offset strokes with longer durations at lower peak velocity. The longer offset durations can be accounted for with the lower peak velocity PWS achieve when transitioning between consonantal closure and the subsequent vowel.

Slower execution has often been claimed to provide more control over execution (Archibald & De Nil, 1999; Nudelman et al., 1987; Tasko et al., 2007). Similar to the overall slowed speech rate observed in the acoustic speech signal (see 3.2.1), the slower and longer kinematic observations of transitions by PWS may suggest increased sensitivity in monitoring (1.4.1.3) – even more so as speakers were informed that the study was designed to investigate their speech. Slowing the speech rate might therefore be a strategy employed by PWS to remain stable and fluent in their productions (Onslow et al., 1992).

In addition to differences in kinematic duration, results further revealed differences in kinematic variation for the two speaker groups. These differences were most apparent in offset movement strokes, which is also where, according to Wingate's Fault-Line hypothesis, we expected to find differences between speaker groups. Variations of onset durations did not reveal any significant differences for the two groups. For mean offset durations, in contrast, the variation was generally larger for

PWS when compared to PNS, which also reached significance. For offset peak velocity variation, we observed the opposite where PWS performed with lower variation, which however did not reach significance.

5 General Discussion and Conclusion

5.1 Advances of the Current Study

This thesis provides kinematic data testing the Wingate's Fault-Line hypothesis about production differences between speakers who stammer and typical speakers. Further, this thesis is making an important methodological contribution to the analysis of ultrasound tongue imaging (UTI) data. We recorded acoustic and articulatory data of CV utterances produced by PWS and PNS.

Fluent productions of people who stammer were established using an objective auditory perceptual judgement task. The judgement is confirmed through acoustic measures, which we will lay out in the following section.

5.1.1 Acoustic Measures vs. Perceptual Salience

Prior to investigating potential differences between speaker groups, we explored the validity of the measures. Acoustic measures appear to be valid for three reasons:

- Duration measures show shorter schwa segments when compared to /ɑ/ and /i/ vowels, as would be expected considering that schwa is typically reduced in contrast to corner vowels like /ɑ/ and /i/ (Browman & Goldstein, 1992b).
- Locus equations were sensitive to differences in coarticulation as a function of the consonant. The pattern is in line with the framework of the degree of articulatory constraint model of coarticulation (Recasens & Espinosa, 2009) where tongue body involvement increases coarticulatory resistance. Labial consonants like /p/ are therefore expected to show largest degree of coarticulation, which is represented with the steepest locus equation slope as could be observed in our data.

- Formant slope measures exhibit a clear relation between formant slope duration and formant slope extent. Longer durations imply that greater distances are covered. This aspect follows the rate-time-distance equation as defined in traditional physics. Our data follow that equation despite the non-steady motion including acceleration and deceleration.

Seeing that these differences show as clearly as they do, they underpin the validity of differences we found in the fluent data of the two speaker groups. Perceptually, the two speaker groups could not be distinguished. The acoustic measures we applied, however, reveal differences for the two groups, which we will lay out in the following.

5.1.1.1 Acoustic Duration & Variation

PWS performed with longer acoustic durations suggesting lower speech rate, which is in line with previous findings (Lenoci, 2018; Wieneke et al., 2001). In contrast to findings by Archibald and De Nil (Archibald & De Nil, 1999), the results of the current study do not provide evidence that severity of the stammer has explanatory value for movement speed. Slower execution can be explained with increased monitoring of the speech output to maintain fluency. The PWS anticipates that he or she may become disfluent and slows down speed to maintain full control and thereby fluency.

With regard to variation, we observed significantly larger variation in PWS than in PNS, particularly in the case of closure durations. Variation is often interpreted as a measure of control where increased variation is considered equivalent to less consistent movement execution, typically related to diminished control (Barbier, Perrier, Ménard, Payan, et al., 2013). Seeing that only perceptually fluent data was included in the analysis, these results suggest that the acoustic measures capture subtle indicators of disfluency in the perceptually fluent speech of PWS. The largest difference in variation is observed for closure durations in PWS. Closure durations are also greatly affected in stammered blocks, where the tongue is in extended contact with the roof of the mouth (palate) and release is delayed (5.3.1). The larger

variation for closure duration may therefore suggest that remainders of perceptually non-salient disfluencies, such as blocks, are included in the data, something we will come back to later in the discussion.

5.1.1.2 Acoustic Coordination

Further, results revealed flatter locus equations and greater formant slope extents for PWS, which implies that PWS coarticulate less when compared to PNS. Locus equations are regression lines that inform about the degree of anticipatory coarticulation a consonant received in relation to a number of vowels. The steeper the locus equation slope, the greater the overall coarticulation of that speaker – or speaker group. The flatter locus equation slope found for PWS is echoed by the greater formant slope extent. Formant slope extent describes the change in F2 value when transitioning from one segment to the next. Overlap of these adjacent segments would draw the F2 values of these segments closer to one another, also referred to as coarticulation. In contrast, lesser overlap of adjacent segments results in a decreased degree of coarticulation where adjacent segments maintain their inherent quality, thereby maintaining their formant structure. Seeing that coarticulation is a means to articulate strings of sounds more efficiently (Fowler, 1980; Guenther & Guenther, 1995; Saltzman & Kelso, 1987), the lack of coarticulation for PWS must be more costly for articulatory movement control. This may also account for the longer durations observed in PWS.

All of these acoustic differences support the Fault-Line hypothesis as they all target the transition between the consonant and the subsequent vowel and indicate that PWS perform the speech task more slowly and more variably than PNS. The differences mentioned were captured using acoustic measures, all of which provided statistically significant findings for comparisons between speaker groups. Although identifiable through statistical analysis of acoustic measures, the differences could not be detected perceptually. This highlights the difference between perceptual and acoustic salience and emphasises the value gained from acoustic analysis where acoustic detail enables a deeper understanding.

5.1.2 Methodological Advancement

In this section we will briefly present the advances of the applied kinematic methodology and explore the validity of the results obtained. We will further evaluate any additional value of these measures over and above the different types of acoustic measures presented (see Chapter 3).

Articulatory data is the most direct source to gain insights into motor impairments, such as stammering. Acoustic analysis is typically based on waveforms that combine the entire articulatory information in one channel. When compared to acoustics, articulatory analysis provides a more direct access to the speaker's system of control. The additional articulatory information may therefore lay the ground for new perspectives in the investigation of motor control.

In this thesis, we applied a method to obtain dynamic information from kinematic articulatory data that combines the benefits from two approaches to analysing articulatory data: ultrasound tongue imaging (UTI) and electromagnetic articulography (EMA).

UTI is a non-invasive articulatory instrument (Cleland et al., 2016) providing midsagittal information of the tongue surface moving in time and space (1.5.3.3). Further, the instrument is both accessible and cost efficient. With high-speed ultrasound it is possible to capture articulatory kinematics. The standard approach in UTI has been almost exclusively concerned with the static analysis of tongue shapes, or at best comparing small numbers of consecutive (sparsely sampled) static images. This was in part due to the relatively low frame rate of video ultrasound and is now no longer necessarily appropriate for data collected using high-speed systems. To compare data within and across speakers, referents for data orientation are required. Previously, attempts were made to define external referents that would allow for data orientation (Zharkova, 2013) based on, for example, the real space or vocal tract size. External referents like vocal tract sizes, however, differ between individuals (Fant, 1966; Fitch & Giedd, 1999; Kent & Moll,

1969). Further, even within individuals, external referents may change when the individual changes posture (Zharkova et al., 2015). For a single recording session, this can be controlled using a headset to stabilise the ultrasound probe (Articulate Instruments Ltd, 2008; Scobbie, Wrench, & Van Der Linden, 2008). The lack of a data-intrinsic referent therefore makes it difficult to compare intra- and interspeaker differences. The lower number of UTI frames and the lack of a referent for interspeaker comparison would typically provide for qualitative analyses of UTI data, suited for accessible mal-articulations (Cleland et al., 2015; Heyde et al., 2017). To capture differences in articulation that are not as easily accessible, however, quantitative analysis of larger datasets is required.

EMA is the current gold standard for quantitative articulatory research, as it offers high-resolution data with only minor errors in precision while using the 3D space of the speaker. It provides insights into lingual kinematics for any sound (independent from place or manner of articulation), as well as for the transition between sounds. EMA, however, is limited concerning the nature of the data collected, the tolerance of participants and the accessibility of the instrument.

Combining the advantages of UTI and EMA we applied a novel approach to the analysis of articulatory data, using a data-intrinsic vector. The vector was employed to control for data orientation allowing for quantitative within- and across-speaker comparisons. We recorded high-speed UTI, which allows investigation of most of the tongue surface contour moving in time. The large number of ultrasound frames provided material for dynamic analysis. Tongue contour kinematics informed the location of a vector along which measures were obtained. For each speaker and recording session, the vector was placed anew, which accounts for individual differences of speakers and the slightly different probe placement for each session. The vector was located where maximal lingual displacement could be captured along the vector. Because the measurement vector is based on tongue kinematics, it is considered intrinsic to the data. Using a measurement vector that is intrinsic to

the data, the applied method accounts for differences within and across speakers – thereby rendering the method replicable.

Measures were adopted from traditional EMA measures where velocity profiles inform about duration and peak velocity of a trajectory. Measures represent trajectories in transition from consonant to the subsequent vowel. In particular, we observed two movement strokes – onset strokes into the closure and offset strokes from the closure to the subsequent vowel. The approach applied in this thesis is in many respects similar to traditional dynamic analyses of EMA data, which involve duration and velocity measurements for trajectories of pellets attached to the various points on the tongue and other articulators. It is therefore fully appropriate for a study of kinematic aspects of the tongue movement. We applied these measures to test the Fault-Line hypothesis on fluent speech of people who stammer. Wingate's Fault-Line hypothesis suggests that PWS struggle in transition, which we expected to observe for offset strokes, while no difference between groups would be expected for onset strokes. We obtained quantitative kinematic measures. The high-speed ultrasound provided a large dataset with sufficient power to perform statistical testing. Using statistical analysis, even small differences could be captured that were not be accessible using qualitative analysis.

5.1.3 Articulatory Measures vs. Acoustic Measures

The kinematic measures show a relation of duration and velocity. Though we cannot assume consistent movement, we would still expect to find proportionally longer durations as peak velocity decreases. Our measures confirm that longer durations can be related to lower peak velocity values. Onset strokes were shorter in duration and produced at higher peak velocity than offset strokes. This was observed for both speaker groups equally.

Articulatory measures reflect phonetic differences, with differences in material design being captured. Onset strokes represent the transition between prothetic schwa and velar closure for all recordings that were analysed articulatorily. Because

the targets at either end of the onset stroke transition were stable, we did not expect any differences as a function of vowel in onset stroke measures. This was supported by the findings. Offset strokes in comparison included differing phonetic contexts. They represent the transition between velar closure and the subsequent vowel. Seeing that the quality of the three target vowels differed, differences in measures were expected. Starting from the velar closure where the tongue body is elevated, the transition to higher vowels would be shorter and require less time than transitions to lower vowels. This was also confirmed by the results showing a clear durational difference between the high and the mid and low vowels. Making these details accessible, confirms the measures' validity.

5.1.3.1 Group Differences

Comparing onset and offset strokes, offset strokes are found to be naturally longer, reaching lower peak velocity. Including group as a factor, statistical analysis revealed an interaction for trajectory and group where PWS produce significantly longer and more variable offset strokes when compared to PNS. Their offset strokes were moreover produced at significantly lower peak velocity. As mentioned earlier, offset strokes represent the transition from the consonant to the subsequent vowel. Group differences in offset stroke duration and velocity therefore are in keeping with Wingate's claim that PWS struggle in transition. Further support for the claim that PWS struggle in CV transition, comes from the lack of a difference in onset trajectory. Here, speaker groups behave comparably on both duration and peak velocity measures.

5.1.3.2 Peak Velocity & Duration

The longer durations and lower peak velocity for PWS in offset strokes suggest overall less refined movement control when compared to PNS.

The longer articulatory durations echo the longer acoustic durations we observed for PWS (see section 3.2.1). In combination with the overall lower peak velocities for PWS, these articulatory measures appear sensible and legitimate to base our group comparison on.

Differences in duration may result from difficulty in coordination, which is in line with the greater formant slope extent and the lower locus equation slope for PWS. The latter two are established measures of coarticulation. Typically, coarticulation causes adjacent segments to be produced in overlap, which entails that individual segments adapt to the formant structure of their adjacent segments. The larger formant extent for PWS, however, reveals that adjacent segments tend to maintain their inherent formant structure. They do not adapt to the formant structure of adjacent segments to the same degree as they do for PNS. Larger formant slope extent for PWS may therefore be an indicator for less coarticulation.

If it is the case that PWS coarticulate less, they may struggle with the joining together of segments and more particularly with the trajectory between them. While coarticulation makes transitions efficient, without coarticulation the production of more than one segment is likely to become inefficient, possibly yielding an increase in articulatory effort. In this context, the longer durations and the lower peak velocity may be indicative of the added effort when joining adjacent segments. The slowing of speech may be both reactive and proactive in that it makes articulation more manageable. This is in line with claims made by Max and colleagues (Max et al., 2004) that in PWS feedforward is weaker and an overreliance on feedback control may require too many resources to simultaneously also move forward in the production of speech. This may inhibit the speed at which speech is produced.

5.1.3.3 Variation

The difficulty in transitioning between segments is further supported by the larger variation found for PWS. Offset durations are shown to be more variable in PWS. The increased variation shows consistently for each vowel context. The larger variation for PWS was also found for acoustic release and acoustic closure durations further supporting the claim of less refined motor control with difficulty particularly in consonant-vowel transition.

There may be two explanations for the increase in variability: On one side, the PWS data may consist of generally less consistent productions indicating an overall less stable motor command of PWS. On the other hand, it may be possible that the increased variation in the PWS data is driven by only few data points which may be remainders of disfluencies. The perceptual categorisation of data into fluent and disfluent categories may have missed few instances of disfluencies, now responsible for the larger variation. The latter view suggests that PWS data as binary – either fluent or disfluent, while the former sees PWS productions as situated on a continuum of fluency.

We explored the plausibility of these two options and investigated the distribution of the offset durations. While the homogeneity of variance returns a significant difference for the two speaker groups, this does not establish whether offset durations are binary (fluent – disfluent) or globally more variable. In order to see in what way these two distributions differ, we visually examined the data. Visual examination showed that durations in the PWS data were overall more variable, supporting the notion that the transition in PWS is globally less stable when compared to PWS.

Comparing the kinematic findings to findings from different types of acoustic measures, we found that the method applied can be useful not only in validating acoustic findings, but further in making available articulatory details not captured in the acoustic signal (see section 5.3).

5.1.4 The Fault-Line

Despite a widely-held belief that data can be categorised into fluent and disfluent productions, the acoustic and articulatory analyses show that even perceptually fluent data from PWS contains deviant productions. These deviant productions highlight the limitations of perception, for which all data appeared homogenous. Both acoustic and articulatory analysis, found differences between the perceptually fluent data from PWS and data from PNS.

Having found differences in the fluent PWS data, has two main implications:

- The measures applied are sensitive enough to capture differences that are perceptually non-salient.

- Stammering affects speech more globally than the local events which are perceptually identifiable and therefore typically classed as instances of stammering (consistent with earlier research, e.g., van Lieshout, Hulstijn, & Peters, 1996; Ward, 1997).

The latter point highlights the importance to distinguish between the disorder of stammering and the perceptual available moments of stammering – as implied by Wingate in his Fault-Line hypothesis (Wingate, 1988; see section 1.4.2.1). Moments of stammering are usually used to inform about the presence or absence as well as the severity of the disorder of stammering. Perceptually available information does, however, not provide a holistic impression where symptoms of the stammer may be present even when not perceptually available.

While differences between the two speaker groups were found, they did not show on each measure. The reason why not every measure returned differences for the two speaker groups can be related to the fact that only the perceptually fluent productions of both speaker groups were analysed where differences may be more subtle. Instrumental techniques in addition to perceptual information can provide a more in-depth understanding of the disorder of stammering.

5.1.4.1 The Fault-Line vs. Coupled Oscillator Model

Differences distinguishing the two speaker groups are most prominent in acoustic closure duration variation, as well as in articulatory offset stroke duration and variation, all representing different perspectives to the consonant-vowel transition.

The coupled oscillator model distinguishes in-phase and anti-phase, where offset strokes fall into the category of in-phase. The model predicts that in-phase co-ordination is generally easier compared to anti-phase co-ordination. The Fault-Line

hypothesis, in contrast, predicts that consonant-vowel transitions (corresponding to in-phase as defined by the coupled oscillator model) are more difficult to manage due to the tight coupling of the two gestures.

Our results show that PWS do not struggle to reach the velar closure (anti-phase). Instead, PWS show longer and more variable offset strokes, indicating difficulty in CV transition (in-phase) thereby:

- contradicting difficulty in in-phase (as predicted by the oscillator framework used by AP).
- providing support for a transition deficit in PWS (as predicted by Fault-Line hypothesis)

The transition deficit proposed by Wingate explains the differences in offset strokes observed in the perceptually fluent speech of PWS and PNS. What it does not explain, however, are the disfluencies that occur in transition to the consonant (in-phase) as presented in section 5.3.3.

closure durations in anticipation …

5.1.4.2 The Fault-Line vs. DIVA Model

Another possible explanation for the longer offset durations as well as the larger closure duration variation observed in the articulatory data may be difficulty with consonant coordination more generally.

The articulatory analysis was performed for velar plosive consonants, where it should be acknowledged that plosive consonants consist of several phases. Once maximal closure is achieved, pressure needs to be build up to prepare for the release phase. Problems with the release phase of the consonant may account for both longer closure durations as well as longer offset durations.

Difficulty with consonant coordination would be in line with the hypotheses proposed by Max and colleagues (Max et al., 2004) who applied the DIVA model

(1.4.2.2; Guenther, 1994) to stammered speech. The DIVA model uses feedback and feedforward control mechanisms. Children rely on feedback to map it to the articulatory information. This feeds into the feedforward control. With increasing feedforward control, targets become narrower.

In PWS mapping may be instable and targets too narrow which is potentially causing difficulty at the consonant. The overreliance on feedback mechanisms may result in longer durations in PWS. The narrow targets may leave less flexibility for PWS, which may account for the decreased variability found for formant slopes and locus equations.

In this case the differences in transition would be residual to the difficulty at the consonant, thereby:

- providing support for instable mapping in PWS (as predicted by DIVA)
- providing support for overreliance on feedback control (as predicted by DIVA)
- providing support for a transition deficit in PWS (as predicted by Fault-Line hypothesis)

Instable mapping and overreliance on feedback control may also account for instances of disfluencies similar to those presented in sections 5.3.

5.2 Limitations of the Current Study

In the following section, we will briefly cover some limitation to the study. We will begin with the methodological limitations and then cover two aspects related to the quality of the data.

5.2.1 Methodological Limitations

We have obtained acoustic and articulatory measures of the fluent speech of PWS and PNS. More particularly, lingual movements for /ə/ + CV utterances were investigated. This thesis focuses on articulations that can be captured clearly via ultrasound imaging in the sagittal plane and therefore findings cannot be generalised to other productions (e.g., laterals). The lingual movements were broken down into two main movement stokes that were quantified employing measures of duration, maximum velocity, and average speed. The first movement stroke was that of the tongue approaching the palate for consonantal closure, i.e., onset. The second lingual movement stroke was that of the tongue moving out of / away from consonantal closure into a relatively stable articulatory position in the subsequent vowel.

5.2.1.1 Tongue Advancement when moving towards /i/

The materials in this study included symmetrical (ə _ɑ and ə _ə) as well as asymmetrical (ə _i) contexts. Less symmetrical contexts need to take into account loop-like trajectories (Mooshammer et al., 1995) and may therefore not be directly comparable.

Looking at duration and peak velocity differences in vowels, this tendency for larger durations at lower peak velocity and vice versa does not hold for all instances. We observe, for example, lower duration values for /ki/ compared to /kɑ/ and /kə/. The shorter durations for /ki/ do not come with higher peak velocity values. Instead, /ki/ shows low peak velocity values compared to /kɑ/ and /kə/.

The combination of lower duration and lower peak velocity values for /ki/ compared to /kɑ/ and /kə/ (especially in offset trajectories) can be explained by how the tongue surface movement along the measurement vector is captured. Because the measurement vector captures vertical movements from the origin of the ultrasound probe along the vector only, advancing tongue movements in the vocal tract are not captured. The difference between /i/ as compared to /ɑ/ and /ə/

can be explained by the tongue looping through space – involving vertical as well as horizontal movement (Birkholz, Hoole, Kröger, & Neuschaefer-Rube, 2011). The measures therefore reflect the more advanced tongue movement required for the transition (offset trajectory) into the vowel /i/ compared to the larger vertical movement that can be captured along the measurement vector for the lower vowels /ɑ/ and /ə/. While this circumstance might render direct comparisons of transitions with different degrees of tongue advancement less powerful, it does not affect the comparison between speaker groups.

5.2.1.2 Measurement Vector & Amount of Overlaid Splines

Using a single vector to obtain measures means that the original data is reduced by a dimension. Typically, ultrasound images consist of black and white pixels in a two-dimensional space. With multiple ultrasound images showing the tongue position at different points in time, a third temporal dimension becomes available. For the present study, the two-dimensional ultrasound image was reduced to a vector that was located along a single scanline.

The fact that the data is reduced by a dimension as measures are taken along a vector highlights the importance of that vector. The vector was located along a scanline along which the maximal movement could be observed over time. The extent to which there was movement and the location where there was most movement were based on overlaid splines of the V_PCV sequence.

It remains open to investigation to what extent a different number of overlaid splines (see Figure 30) would result in the same scanline to locate the measurement vector. Further, it is uncertain to what extent the measures would be affected if, for example, the vector was located on a neighbouring scanline (see Figure 31). Hence, using a more intuitive approach could possibly produce results that are equally meaningful. As such, establishing the location of a vector through mere eyeballing of the important area on the tongue surface is surely a less reproducible but also less time-consuming method.

5.2.1.3 Smoothing Displacement Curves

What is critical about smoothing is that the right balance needs to be found. Data needs to be smoothed to remove artefacts. But too much smoothing can remove information that may be relevant to the participant of interest. For the current study, for example, too much smoothing would have meant to remove intended tongue movements along the measurement vector.

Seeing that tongue movement is ballistic in nature, we can expect a single curve for each displacement. We determined a smoothing factor that was the same across all speakers. Probing with data of control speakers first, the factor was adjusted to obtain a most clear curve representing the displacement data of the tongue moving along the measurement vector. As was expected we would see the first half of a curve representing increasing elevation of the tongue coinciding with the tongue moving towards the palate for consonantal closure at /k/. The second half of the curve would then show the lowering of the tongue as it moves away from the palate and into the vowel. To establish the final smoothing factor in this study, it was tested against the control data. A protocol for determining the 'correct' smoothing factor, however, is yet to be established as no protocols for data smoothing in ultrasound are in place.

5.2.1.4 Closure Threshold

Another critical process takes place in defining the segments to be measured. Displacement curves provide a relatively clear picture of the displacement of the tongue body moving along the measurement vector. Beginning and end of these curves are challenging as these indicate that the tongue starts off from a stable position and ends at a stable position with close to zero displacement. Zero displacement is highly unlikely taking unintended lingual movement into account. To circumvent this issue, we adapted a technique typically employed with data from electromagnetic articulography. A threshold of 20% of the maximum displacement was taken to flag the beginning and end of directed /intended movement. Though

the 20% threshold is an established value in EMA studies (Kühnert et al., 2006), no alternative thresholds were tested for ultrasound data.

5.2.2 Speech Rate

As was shown in both the acoustic (3.2.1) as well as the articulatory measures (4.2.1), PWS performed with longer durations when compared to PNS. As previously discussed (3.3; 4.3) the longer duration measures suggest overall lower speech rate in PWS when compared to control speakers. Lower speech rate was also observed in previous studies in which the slowed speech was associated with a decreased demand resulting in increased control and thereby enhanced fluency in PWS (Andrade et al., 2003; Andrews et al., 1982; Archibald & De Nil, 1999; Guitar, 2013; Nudelman et al., 1987; Onslow et al., 1992; Parks, 2001; Postma et al., 1990; Tasko et al., 2007). The longer durations could be a result from the transition deficit proposed by Wingate. They could, however, also result from a strategy applied by PWS to maintain fluency.

Aiming at the speakers' most habitual speech, we have not instructed participants to speak at a certain rate. For the same reason we did not instruct the PWS to employ / not employ strategies acquired during speech therapy.

It is therefore not clear whether PWS performed at their habitual rate. But it is also difficult to state what would be expected as habitual rate for an experimental situation. While PWS were not instructed to change their speech rate, it needs to be acknowledged that this is an experimental situation and PWS might have adjusted their speech rate involuntarily. Future studies may want to obtain conversational or read speech material for direct comparison.

5.2.3 Speaker Variation

In the coming sub-sections, we will briefly discuss potential limitations with regard to the quality of data used for the investigation of fluency. What makes it difficult to

study fluency, is the variability it presents with. There are multiple factors that affect fluency in a speaker's speech. Speech varies in response to the speaker's physical (e.g., age, height, weight) as well as their psychological state (e.g., emotions like anger, fear, and sadness). Moreover, speakers vary in their responses to similar situations. Thus, one speaker might be more fluent in relaxed social situations, while stressful situations such as public speaking may cause that same speaker to be more disfluent, while for someone else, a formal presentation in front of a big audience might motivate a performance with speech far more fluent than occurs in natural intimate conversations (Jackson, Tiede, Beal, & Whalen, 2016). In addition, even if speakers have similar responses to context, one speaker may present with more extreme or more frequent disfluencies than another.

Given the natural variation in the fluency of speech, we see that disfluent speech is not necessarily an indicator of pathological speech. Disfluencies are part of every typical speaker's repertoire (which may even be useful functionally as involuntary, or even voluntary means to hold the conversational floor). Pathological disfluencies in contrast are primarily involuntary, can be severe, frequent, and functionally debilitating. The question is how to define them – especially since speech is highly variable depending on the physical and psychological state of each speaker. The high variability of speech automatically leads to fuzzier boundaries between the categories of fluent and disfluent speech and makes it even more difficult to distinguish them.

5.2.4 Fluency Judgement

Data produced by PWS were categorised as perceptually fluent and disfluent. Only data perceptually categorised as fluent was included in the subsequent analysis to explore whether even the fluent speech of PWS deviates from that of typical speech.

We employed a two-stage fluency judgement approach to categorise the recordings into the two categories of fluent and disfluent recordings. Following acoustic and visual inspection of the ultrasound data, listeners were asked to categorise the pre-

selected data into fluent and disfluent recordings. The disfluent category included recordings where the majority of listeners agreed with certainty that these recordings are disfluent. All other recordings were subsumed in the fluent category. From the previously established 25 deviant cases, 7 recordings from 3 speakers (5 of which come from just one speaker) were recognised as overt disfluencies (see Table 41) and the remaining 18 recordings were recognised as covert disfluencies and uncategorised behaviours.

Table 41 *'overtly' disfluent recordings with acoustic segmental durations (ms)*

Speaker	Prompt	Session	closure	release
5	kɑ	1	536.4	72.3
5	kə	1	106.1	77.1
5	kə	1	892.3	190.9
5	kə	1	1596.4	147.8
5	kə	2	43.9	83.2
6	kɑ	1	118.6	86.2
10	kɑ	1	95.4	54

Because we employed a perceptual judgement task in the second stage, the categories were compiled based on overt fluency / disfluency. The term disfluency in the context of the perceptual judgment task (in contrast to the understanding of disfluency as incorporating overt and covert phenomena throughout the thesis) refers to overt disfluency exclusively. By overt disfluency I refer to disfluency that is marked by acoustically salient disruptions in the speech flow. Covert disfluencies in contrast to overt disfluencies are not acoustically salient. In a perceptual judgement task, covert disfluencies may therefore remain uncovered and are most likely assigned to the category of fluent speech. This categorisation of fluent and disfluent speech, in fact, resembles the way that speech of PWS is perceived in everyday life, i.e., based on perception. When we consult articulatory observations, however, this seemingly clear-cut distinction is supplanted by a different understanding of what

disfluency is (i.e., where we need to distinguish perceptually available moments of stammer and the disorder of stammering which may present with symptoms that are not necessarily perceptually available).

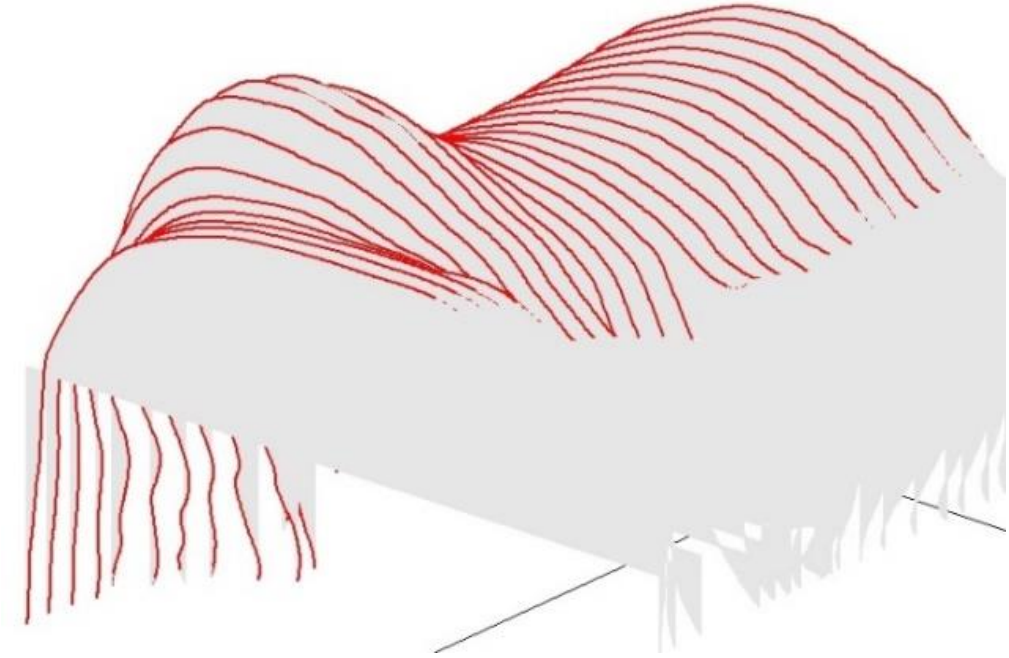


Figure 33 Tongue splines (root to tip) for a perceptually fluent production of /ə kɑ/ over time (front to back)

In Figure 33 we can see the splines for the tongue surface as they move through /ə kɑ/. The tongue surface begins relatively stable at the schwa (front), moves into velar closure for /k/ and transitions on to the subsequent vowel /ɑ/. This sequence is executed smoothly, which is also captured in the fluent perception. Acoustics and articulation, however, do not necessarily match.

There are motor anomalies which are not perceptually salient, but are consistent with stammering. While these anomalies may not necessarily be perceptually salient, they can be observed on the acoustic and underlying articulatory level. Disfluencies are generally understood as perceptually salient breakdowns in the fluency of speech that can be observed on the surface. There are, however, covert

disfluencies that are not perceptually salient as PWS manage to conceal them. They are, however, still part of the disorder of stammer. These covert disfluencies (together with the overt disfluencies) are observable on the underlying articulatory level for which the categorisation might be more complex than that for overt disfluencies (i.e., block, repetition, and prolongation).

Following the perception-based fluency judgement conducted for the current study, covert disfluencies would fall into the category of fluent productions. The inclusion of perceptually non-salient disfluencies in the analysis of perceptually fluent productions may therefore account for the significant differences in variation observed for PWS. The complexity of this topic becomes more valid when we consider the nature of disfluencies, which we will briefly discuss in the following section.

5.3 Disfluent Recordings

While we have an idea of what disfluencies sound like, the underlying articulatory level has not received much attention (see Figure 34). Three very distinct articulatory observations of disfluencies will be discussed. Within the recordings that were found to be deviant we could observe three very distinct articulatory patterns which will be described below. While the first pattern appears to map in some ways onto the acoustic description of blocked disfluencies, the latter show articulatory patterns that do not correspond to the traditional categories of overt disfluencies (i.e., repetitions, prolongations, and blocks).

5.3.1 Continued Contraction

The pattern that was easiest to identify was that of the tongue in extended contact with the palate during consonantal closure (Figure 34). For velar consonants, we could observe the following: The tongue body raises, tongue tip and tongue root contract and the overall elevated and contracted tongue configuration is

maintained for noticeably longer than would be expected before initiating the transition in fluent speech.

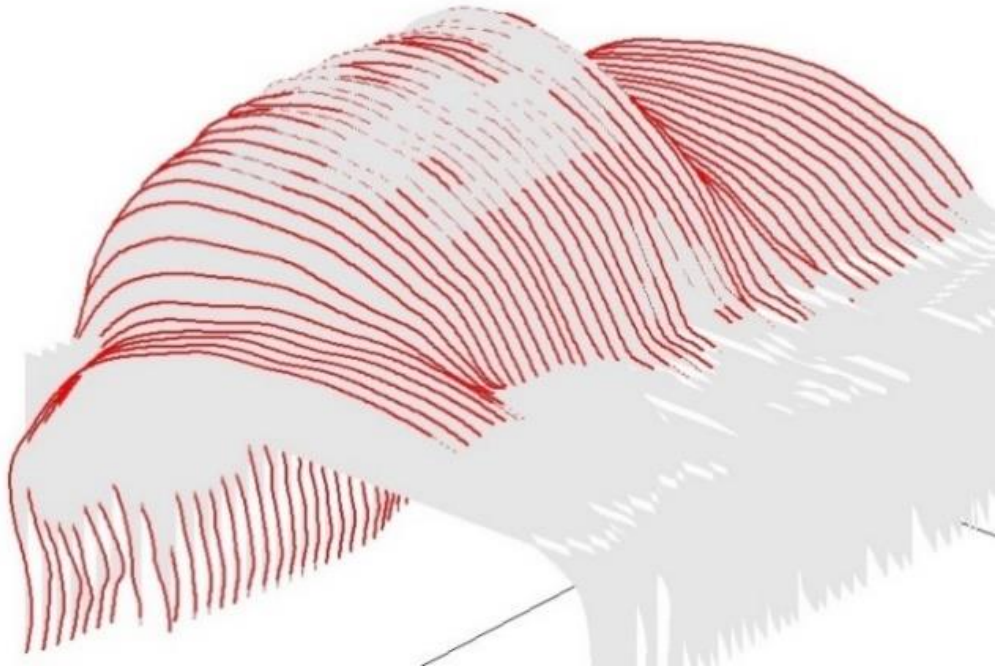


Figure 34 Tongue splines (root to tip) for a perceptually disfluent production of /ə kɑ/ over time (front to back)

Acoustically, this appears as an extended silent closure. What we learn from the ultrasound recording is that the tongue body remains motionless in palatal contact during the closure. Other parts of the tongue, i.e., tongue tip and tongue root, however, continue to contract until a point where the entire tongue relaxes, tongue tip and root release, and the tongue body begins to move away from the palate into the release. The overall prolonged motionless posture at consonantal constriction is accounted for with the term block. The continuing constriction of anterior and posterior parts of the tongue throughout the extended blocking phase, however, is a characteristic that we can only observe in the articulatory data.

We observe the successful achievement of the target where the tongue reaches maximal closure. After achieving the consonantal closure, plosives consist of a second stage where intraoral pressure needs to be built up and coordinated to enter the release phase. The delayed transition raises the question whether the PWS struggles indeed when transitioning or rather on the consonant when initiating the transition. In light of Wingate's Fault-Line Hypothesis, it can be argued that the consonant cannot be released until the vowel is ready for execution. This would account for the longer closure durations observed for the disfluent recordings (see Table 41) which is also in line with the longer offset durations observed in the articulatory data for PWS.

5.3.2 Double Bumping

A second pattern that occurred a few times in the disfluent kinematic data of PWS speakers (PWS 5 and PWS 6) is that of a double bump at consonantal constriction. The tongue body raises to reach the target at closure and then slightly moves away from the target before approaching the target again and moving into the release phase. The slight movement out of the closure might indicate a very short relaxation phase surrounded by two movement phases into and out of the target.

Typically, velocity transitions are smooth with a single curve for onset and offset trajectories. Velocity decreases when approaching consonantal closure and increases again around the release of intraoral pressure as the tongue moves away from the constriction. The smooth transition from velocity decrease to increase suggests that kinetic energy is employed efficiently without inserting superfluous movements that could interrupt the generally smooth motion. In the disfluent data, however, we observe superfluous movement resulting in a double bump at consonantal closure.

The additional motion observed in the disfluent data could be accounted for by two explanations: First, a surplus of energy and second, a timing issue: The double bump could indicate an overshoot at palatal constriction resulting from a surplus of kinetic

energy (Ballistic Tongues: Gick, Wilson, & Derrick, 2012 chapter 9.2.1) or inaccurate mapping of the target at the palate. It may well be that in contrast to fluent smooth productions, the amount of kinetic energy that is used to approach consonantal closure is not as well adjusted. Energy is required for the tongue to reach the palate. In cases where a double bump can be observed, a surplus in energy could be responsible. More kinetic energy requires equally more force to be stopped. This is where we see the tongue bouncing off the palate.

Another explanation that could account for the additional movement is the lack in temporal coordination during the closure phase. While the target is achieved, the speaker appears to struggle when initiating transition to the subsequent vowel. Together with the successful consonantal closure, speakers need to build up intra-oral pressure as the second phase of the plosive. Once enough pressure has built up, the release of the consonant is initiated. It may be that in cases where the tongue bounces off the palate, it is mistimed with the build-up of intra-oral pressure which is why not enough intra-oral pressure can be built up in time for a smooth transition into the release phase. Having to restart at the palate in order to build up intra-oral pressure for the following consonantal release could cause the tongue to bounce off at the palate. This account suggests the transition deficit proposed by Wingate (Wingate, 1988) to be residual to faulty execution of the second phase of the consonant which moreover aligns with the explanation provided for the continued contraction (see 5.3.1)

5.3.3 Inaudible Groping

A third pattern was striking with respect to the complexity and duration (Fig. 36). We could observe extensive covert groping behaviours before the participant was able to reach the consonantal target to initiate the CV target syllable. In one recording the participant shows three attempts to reach the target before successfully reaching the consonantal closure after 400ms. The entire length of the

disfluency was inaudible and only became salient using instrumentation that renders articulations visible (Liss & Weismer, 1992).

This pattern occurred when the speaker was cued to produce an utterance. The cue was followed by the tongue moving into the direction of where the consonantal closure would be expected. With each of the three attempts the tongue got closer to the palate but did not reach it. It would probably be misleading to say that the tongue was entirely uncoordinated. Instead, it seemed that we observed an undershoot with each attempt getting nearer the target. This entire sequence was inaudible as no vocalisation was initiated. The speaker appeared to adjust the muscular control and reached the target with the fourth attempt from where the CV transition sounded relatively fluent.

As the movement did not occur in transition, but prior to the achievement of the initial consonant, this is not fully consistent the Fault-Line Hypothesis. The deficit when moving towards a target could be accounted for by narrower targets in PWS. Narrower targets leave less flexibility when moving into / between targets. This would more generally explain struggles that occur before and after targets are achieved – including transition deficits.

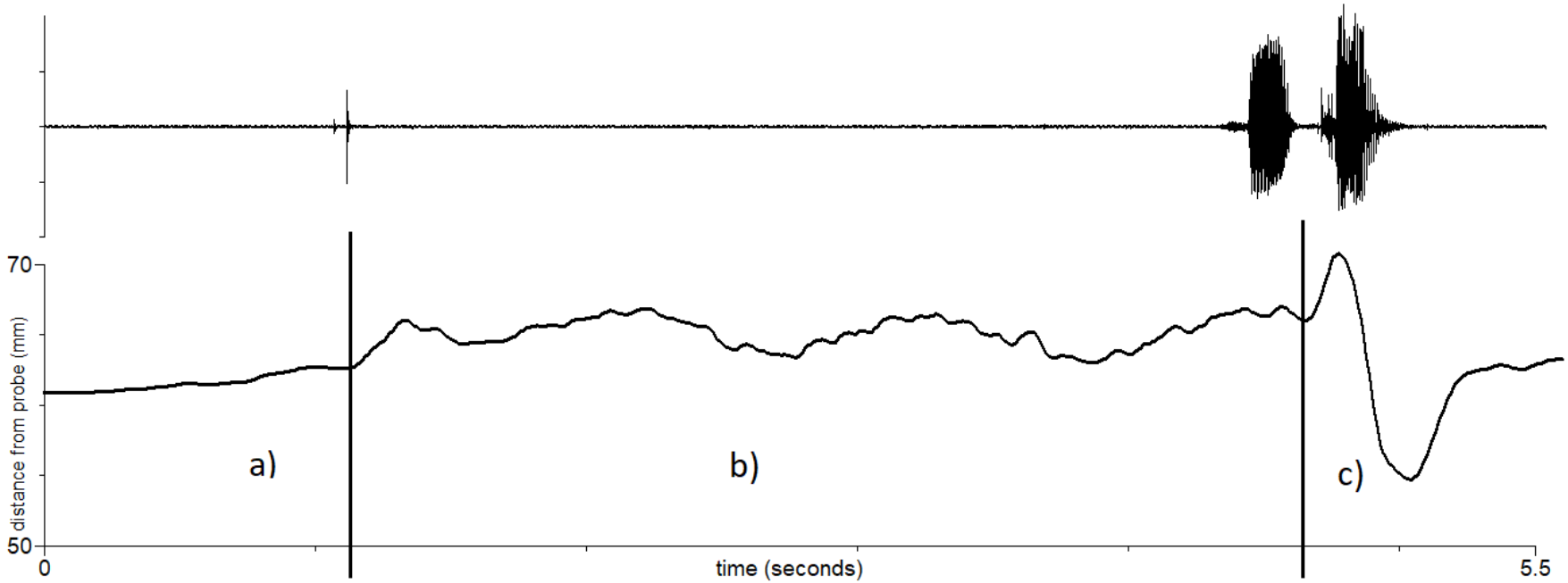


Figure 35 *Acoustic signal (upper panel) and articulatory trace of the tongue displacement for /ə ka/ (lower panel) showing a) still tongue kinematics prior to movement initiation, b) groping behaviour when attempting to reach velar closure, c) successful closure and transition*

5.3.4 Summary

We have observed three very distinct articulatory patterns of disfluency within a very small sample of 25 CV recordings. In a larger data set, one could expect more and perhaps even more distinct articulatory patterns. But even from the three described patterns we can learn something about the relation between articulation and acoustics.

While there appear to be cases where the disfluent articulation can be related to the disfluent acoustic signal (e.g., for blocks), the full complexity of the articulatory pattern is not captured in their acoustic counterpart. For blocks, the prolonged acoustic silence in closure can be directly related to the longer palatal contact before moving into the release. This also matches the description of a block: "*A stutter that is an inappropriate stoppage of the flow of air or voice and often the movement of articulators as well*" (Guitar, 2013). The latter is often perceived acoustically. The continuing contraction of tongue tip and tongue root during the extended closure, however, is not captured in the acoustic signal.

A more extreme example of complex covert articulation is that of inaudible groping where none of the lingual articulation prior to achieving the target at closure was acoustically salient. Trying to relate the groping behaviour to an acoustically established category of disfluency it would probably come closest to the category of repetitions. Repetitions, however, are traditionally defined as repetitions of sounds or syllables, which presupposes initiation of vocalisation. In the observed case, that does not apply. Not sounds, but movement trajectories are repeated.

What these articulatory patterns imply is a complexity and variation that goes beyond that of the acoustic manifestations of disfluency, which typically are categorised in three core types of disfluency, i.e., blocks, repetitions, and prolongations. The increased articulatory complexity and variation raise the question how these articulatory phenomena are related and whether they do have a common underlying cause. The traditional categorisation of core symptoms may

simplify the matter of disfluency to a degree where too much detail gets lost making it impossible to understand what disfluencies are. Further, because only articulatory information can disclose covert disfluencies, both articulatory information should be consulted supplementary to acoustic information when assessing fluency / disfluency.

5.4 Conclusions and Future Implications

Developing and applying a novel kinematic articulatory approach using ultrasound tongue imaging and comparing its findings to established acoustic measures, we were able to show that even the perceptually fluent speech of people who stammer differs when compared to that of typical speakers. Findings revealed differences in both timing and coordination. PWS were overall slower and more variable when compared to typical speakers (PNS). Differences in coordination further suggest that PWS coarticulate less when compared to PNS. Altogether, findings indicate a deficit in consonant-vowel transition that can be observed in the perceptually fluent speech of people who stammer, which is in keeping with Wingate's Fault-Line hypothesis.

The clinical ramifications of these results are twofold: Findings support the use of slide techniques (Fraser, 2007, p. 183; Packman, Onslow, & Doorn, 1994) where PWS are instructed to smoothen and prolong the release of the initial sound and the transition to the subsequent sound targeting CV transitions. Findings may also explain why soft onset techniques may be less effective as they are targeting the consonant and not the transition to the vowel (Stager & Ludlow, 1998).

Comparing the findings from the kinematic approach to that from established acoustic measures allowed us (a) to confirm the validity of the kinematic measure and (b) to explore the value of the acoustic measures, seeing that they are less direct measures of articulation.

The methodology applied to the articulatory data is promising. While in this thesis it was applied to the fluent data of AWS, the same approach may be valuable to the speech of CWS or different speech motor control impairments and also the kinematics of a range of environments in English and other languages. For example, the methodology can be applied to different phonetic contexts. While we have applied it to articulatory data from velar contexts, it is not limited to that. The nature of the measurement vector allows measures in, for example, alveolar contexts. For contexts where the articulatory gesture involves multiple constrictions, the methodology can be adjusted so that each constriction receives a vector along which measurements are taken as shown by Strycharczuk and Scobbie for the case of /l/ darkening (Strycharczuk & Scobbie, 2015).

To further validate the kinematic method applied, it might be useful to collect simultaneous UTI and EMA data as proposed by Aron and colleagues (Aron et al., 2016). Collecting data with both these instruments would allow to compare the vector-based approach to the fleshpoint-based one (Derrick, Best, & Fiasson, 2015).

The design of the materials in the current thesis included a variety of canonical CV syllables that were produced with regular voicing and in whisper. Future studies might want to investigate the behaviour of vowel-initial syllables, which are often initiated with a glottal stop. Future studies may want to investigate how PWS perform in transitions between glottal stop and the subsequent vowel. Do PWS behave comparable in vowel-initial syllables when compared to the material presented in this study?

Two more future approaches concern the voicing modality of the material collected and the experimental setup in which it was collected. Many researchers, including Wingate himself (Wingate, 1969a), have suggested that the mechanisms underlying stammering include difficulty coordinating supra- and super-laryngeal articulation with a particular focus on the transition from voiceless to voiced articulation. Future studies might want to investigate CV transitions in whispered speech or sung speech where voicing is either consistently turned off or consistently turned on as

the switching between voiced and voiceless has previously been stated to be difficult for PWS (Falk, Maslow, Thum, & Hoole, 2016; Healey, Mallard, & Adams, 1976). The findings of the current study support an understanding of stammering as incorporating difficulty at a motor level in transitioning from voiceless to voiced segments.

A tempting future direction would involve the incorporation of laryngeal information via laryngograph. It is important to be aware that such an approach should be used with caution when comparing the speech of PWS to that of PNS as there might be potential confounds from therapeutic approaches targeting the voiced / voiceless contrast.

In light of the findings of the current thesis, it appears that while the perceptual judgement of stammering may be sufficient for many cases, additional acoustic and in particular the articulatory data help understand the kinematics underlying disfluencies in stammered speech.

The work presented in this thesis offers a promising new approach to the use of articulatory analysis in order to explore theoretical understandings of speech pathologies such as stammering.

6 Bibliography

Abercrombie, D. (1979). The accents of standard English in Scotland. In A. J. Aitkean & T. McArthur (Eds.), *Languages of Scotland* (4th ed., pp. 68–84). Edinburgh: W and R Chambers.

Abercrombie, D. (1991). *Fifty years in phonetics*. Edinburgh: Edinburgh University Press.

Adams, D. C. (1999). Methods for shape analysis of landmark data from articulated structures. *Evolutionary Ecology Research*, *1*(8), 959–970.

Adams, M. R. (1987). Voice onsets and segment durations of normal speakers and beginning stutterers. *Journal of Fluency Disorders*, *12*(2), 133–139. https://doi.org/10.1016/0094-730X(87)90019-2

Adams, M. R., & Runyan, C. M. (1981). Stuttering and Fluency: Exclusive Events or Points on a Continuum? *Journal of Fluency Disorders*, *6*, 197–218.

Adams, S. G., Weismer, G., & Kent, R. D. (1993). Speaking rate and speech movement velocity profiles. *Journal of Speech and Hearing Research*, *36*, 41–54.

Alfonso, P. J. (1991). Subject definition and selection criteria for stuttering research in adult subjects. *Haskins Laboratories Status Report on Speech Research*, *105*(106), 231–242.

Alm, P. A. (2004). Stuttering and the basal ganglia circuits: A critical review of possible relations. *Journal of Communication Disorders*, *37*(4), 325–370. https://doi.org/10.1016/j.jcomdis.2004.03.001

Alm, P. A. (2014). Stuttering in relation to anxiety, temperament, and personality: Review and analysis with focus on causality. *Journal of Fluency Disorders*, *40*, 5–21. https://doi.org/10.1016/j.jfludis.2014.01.004

Ambrose, N. G. (2004). Theoretical perspectives on the cause of stuttering. *Contemporary Issues in Communication Science and Disorders*, *31*, 80–91.

Ananthakrishnan, G., & Engwall, O. (2011). Mapping between acoustic and articulatory gestures. *Speech Communication*, *53*(4), 567–589. https://doi.org/10.1016/j.specom.2011.01.009

Anderson, J. D. (2007). Phonological neighborhood and word frequency effects in the stuttered disfluencies of children sho stutter. *Journal of Speech, Language, and Hearing Research*, *50*, 229–248.

Andrade, C. R. F. de, Cervone, L. M., & Sassi, F. C. (2003). Relationship between the stuttering severity index and speech rate. *Sao Paulo Medical Journal*, *121*(2), 81–84. https://doi.org/10.1590/S1516-31802003000200010

Andrews, G., & Harris, M. (1964). *The syndrome of stuttering*. *Clinics in developmental medicine 17*. London: Spastics Society Medical Education and Information Unit, in association with William Heinemann Medical Books.

Andrews, G., Howie, P. M., Dozsa, M., & Guitar, B. E. (1982). Stuttering. Speech pattern characteristics under fluency-inducing conditions. *Journal of Speech, Language, and Hearing Research*, *25*(2), 208–216. https://doi.org/10.1044/jshr.2502.208

Andy, O. J., & Bhatnagar, S. C. (1992). Stuttering acquired from subcortical pathologies and its alleviation from thalamic perturbation. *Brain and Language*, *42*, 385–401.

Arantes, P., Eriksson, A., & Gutzeit, S. (2017). Effect of Language, Speaking Style and Speaker on Long-Term F0 Estimation. In *Proceedings of Interspeech* (pp. 3897–3901). Stockholm.

Archibald, L., & De Nil, L. F. (1999). The relationship between stuttering severity and kinesthetic acuity for jaw movements in adults who stutter. *Journal of Fluency Disorders*, *24*(1), 25–42.

Ardila, A., Ramos, E., & Barrocas, R. (2011). Patterns of stuttering in a Spanish/English bilingual: A case report. *Clinical Linguistics & Phonetics*, *25*(1), 23–36. https://doi.org/10.3109/02699206.2010.510918

Arenas, R. M. (2012). *The role of anticipation and an adaptive monitoring system in stuttering: A theoretical and experimental investigation*. University of Iowa.

Arnold, K. (2015). *Formant frequency transitions in the fluent speech of adults who do and do not stutter: Testing the over-reliance on feedback hypothesis*. Western Michigan University.

Aron, M., Berger, M.-O., Kerrien, E., Wrobel-Dautcourt, B., Potard, B., & Laprie, Y. (2016). Multimodal acquisition of articulatory data: Geometrical and temporal registration. *The Journal of the Acoustical Society of America*, *139*(2), 636–648. https://doi.org/10.1121/1.4940666

Articulate Instruments Ltd. (2008). Ultrasound stabilisation headset users manual: Revision 1.4. Edinburgh.

Articulate Instruments Ltd. (2010). SyncBrightUp users manual: Revision 1.10. Edinburgh.

Articulate Instruments Ltd. (2012). Articulate Assistant user guide: Version 1.18. Edinburgh.

Articulate Instruments Ltd. (2014). Articulate Assistant Advanced ultrasound module user manual: Revision 2.16. Edinburgh.

Articulate Instruments Ltd. (2015). *Tutorial 3: Annotations and splines - How to label your data and fit splines to ultrasonic data*. Retrieved from http://www.articulateinstruments.com/downloads/

Au-Yeung, J., Gomez, I. V., & Howell, P. (2003). Exchange of Disfluency With Age From Function Words to Content Words in Spanish Speakers Who Stutter. *Journal of Speech, Language, and Hearing Research*, *46*(3), 754–765. https://doi.org/10.1044/1092-4388(2003/060)

Austin, W. M. (1941). The Prothetic Vowel in Greek. *Language*, *17*(2), 83. https://doi.org/10.2307/409615

Baayen, R. H. (2008). *Analyzing linguistic data. A practical introduction to statistics*. Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59*(4), 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Babatsouli, E. (2015). Technologies for the study of speech: Review and an application. *Themes in Science and Technology Education*, *8*(1), 17–32.

Badin, P., Tarabalka, Y., Elisei, F., & Bailly, G. (2010). Can you 'read' tongue movements? Evaluation of the contribution of tongue display to speech understanding. *Speech Communication*, *52*(6), 493–503. https://doi.org/10.1016/j.specom.2010.03.002

Bakker, K., & Brutten, G. J. (1990). Speech-Related Reaction Times of Stutterers and Nonstutterers Diagnostic Implications. *Journal of Speech and Hearing Disorders*, *55*(2), 295–299.

Barbier, G., Perrier, P., Ménard, L., Payan, Y., Tiede, M. K., & Perkell, J. S. (2013). Speech planning as an index of speech motor control maturity. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 1278–1282).

Barbier, G., Perrier, P., Ménard, L., Tiede, M., & Perkell, J. S. (2013). Token-to-token variability and anticipatory coarticulation as indicators of maturity of speech motor control in 4-year-old children. In *Proceedings of Meetings on Acoustics ICA2013* (Vol. 19, pp. 1–7). ASA. https://doi.org/10.1121/1.4800623

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R. H. B., Singmann, H., … Green, P. (2015). Package “lme4.” *Convergence*, *12*(1).

Beal, D. S., Gracco, V. L., Brettschneider, J., Kroll, R. M., & De Nil, L. F. (2012). A voxel-based morphometry (VBM) analysis of regional grey and white matter volume abnormalities within the speech production network of children who stutter. *CORTEX*, *49*(8), 2151–2161. https://doi.org/10.1016/j.cortex.2012.08.013

Bell-Berti, F., & Harris, K. S. (1981). A temporal model of speech production. *Phonetica*, *38*, 9–20.

Belmont, A. J. (2015). *Anticipatory coarticulation and stability of speech in typically fluent speakers and people who stutter across the lifespan: An ultrasound study*. *Master Thesis*. University of South Florida.

Ben, G. D. E. L., & Busan, P. (2014). The neural correlates of developmental stuttering : A brief overview of the literature. https://doi.org/10.17469/O2103AISV000021

Birkholz, P., Hoole, P., Kröger, B. J., & Neuschaefer-Rube, C. (2011). Tongue body loops in vowel sequences. In *International Seminar on Speech Production (ISSP)* (Vol. 9, pp. 203–210). Montreal, Canada.

Blackmer, E. R., & Mitton, J. L. (1991). Theories of monitoring and the timing of repairs in spontaneous speech. *Cognition*, *39*(3), 173–194. https://doi.org/10.1016/0010-0277(91)90052-6

Blomgren, M., Robb, M. P., & Chen, Y. (1998). A note on vowel centralization in stuttering and nonstuttering individuals. *Journal of Speech Language and Hearing Research*, *41*(5), 1042. https://doi.org/10.1044/jslhr.4105.1042

Bloodstein, O., & Ratner, N. B. (1969). *A handbook on stuttering*. Delmar Cengage Learning.

Boersma, P., & Weenink, D. (2015). Praat: doing phonetics by computer. Retrieved from http://www.praat.org/

Boey, R. A., Wuyts, F. L., Heyning, P. H. Van De, Bodt, M. S. De, & Heylen, L. (2007). Characteristics of stuttering-like disfluencies in Dutch speaking children, *32*, 310–329. https://doi.org/10.1016/j.jfludis.2007.07.003

Borden, G. J., Kim, D. H., & Spiegler, K. (1987). Acoustics of stop consonant-vowel relationships during fluent and stuttered utterances. *Journal of Fluency Disorders*, *12*(3), 175–184.

Bouchard, K. E., Mesgarani, N., Johnson, K., & Chang, E. F. (2013). Functional organization of human sensorimotor cortex for speech articulation. *Nature*, *495*(7441), 327–332. https://doi.org/10.1038/nature11911

Boucher, V. J. (2008). Intrinsic factors of cyclical motion in speech articulators: Reappraising postulates of serial-ordering in motor-control theories. *Journal of Phonetics*, *36*(2), 295–307. https://doi.org/10.1016/j.wocn.2007.06.002

Braun, A. R., Varga, M., Stager, S., Schulz, G., Selbie, S., Maisog, J. M., … Ludlow, C. L. (1997). Altered patterns of cerebral activity during speech and language production in developmental stuttering. An H2 (15) O positron emission tomography study. *Brain. A Journal of Neurology*, *120*(5), 761–784.

Bressmann, T., Thind, P., Uy, C., Bollig, C., Ralph, W., Irish, J. C., … Bollig, C. (2010). Quantitative three - dimensional ultrasound analysis of tongue protrusion , grooving and symmetry : Data from 12 normal speakers and a partial glossectomee, *9206*(May 2016). https://doi.org/10.1080/02699200500113947

Brocklehurst, P. H. (2008). A review of evidence for the Covert Repair Hypothesis of stuttering. *Contemporary Issues in Communication Science and Disorders*, *35*, 25–43.

Brocklehurst, P. H. (2011). The roles of speech errors , monitoring , and anticipation in the production of normal and stuttered disfluencies.

Brocklehurst, P. H. (2013). Stuttering prevalence, incidence and recovery rates depend on how we define it: Comment on Yairi & Ambrose' article epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, *38*(3), 290–293. https://doi.org/10.1016/j.jfludis.2013.01.002

Brocklehurst, P. H., Drake, E. K. E., & Corley, M. (2006). Perfectionism and stuttering: Findings of an online survey. *Journal of Fluency Disorders*, *44*, 46–62. https://doi.org/10.1016/j.jfludis.2015.02.002

Brooks, P. J., & MacWhinney, B. (2000). Phonological priming in children' s picture naming. *Journal of Child Language*, *27*(2), 335–366.

Browman, C. P., & Goldstein, L. (1989). Articulatory gestures as phonologica units. *Phonology*, *6*(SR-99), 69–101. https://doi.org/10.1017/S0952675700001019

Browman, C. P., & Goldstein, L. (1990). Tiers in articulatory phonology, with some implications for casual speech. *Papers in Laboratory Phonology I: Between the Grammar and Physics of Speech*, 341–376. https://doi.org/10.1017/CBO9780511627736.019

Browman, C. P., & Goldstein, L. (1992a). Articulatory Phonology: An Overview. *Phonetica*, *49*(3–4), 155–180.

Browman, C. P., & Goldstein, L. (1992b). Targetless" schwa: An articulatory analysis. *Papers in Laboratory Phonology II: Gesture, Segment, Prosody*, 26–56.

Browman, C. P., & Goldstein, L. (1995). Dynamics and Articulatory Phonology. *Mind as Motion*, 175–193.

Browman, C. P., Goldstein, L., & Ohala, J. J. (1986). Towards an articulatory phonology. *Phonology Yearbook*, *3*(21), 9–252. https://doi.org/10.1017/S0952675700000658

Brown, R., & McNeill, D. (1966). The "tip of the tongue" phenomenon. *Journal of Verbal Learning and Verbal Behavior*, *5*(4), 325–337. https://doi.org/10.1016/S0022-5371(66)80040-3

Brutten, E. J., & Shoemaker, D. J. (1967). *The modification of stuttering*. Prentice Hall.

Büchel, C., & Sommer, M. (2004). What causes stuttering? *PLoS Biology*, *2*(2), E46. https://doi.org/10.1371/journal.pbio.0020046

Byrd, C. T., Conture, E. G., & Ohde, R. N. (2007). Phonological priming in young children who stutter: Holistic versus incremental processing. *American Journal of Speech-Language Pathology*, *16*(1), 43–53. https://doi.org/10.1044/1058-0360(2007/006)

Carter, P., & Edwards, S. (2004). EPG therapy for children with long-standing speech disorders: predictions and outcomes. *Clinical Linguistics & Phonetics*, *18*(6–8), 359–372. https://doi.org/10.1080/02699200410001703637

Chang, S.-E., Kenney, M. K., Loucks, T. M. J., & Ludlow, C. L. (2009). Brain activation abnormalities during speech and non-speech in stuttering speakers. *NeuroImage*, *46*(1), 201–212. https://doi.org/10.1016/j.neuroimage.2009.01.066

Chang, S.-E., Ohde, R. N., & Conture, E. G. (2002). Coarticulation and formant transition rate in young children who stutter. *Journal of Speech, Language, and Hearing Research*, *45*(August), 676–688. https://doi.org/10.1016/j.jfludis.2011.04.007

Chang, S.-E., Zhu, D. C., Choo, A. L., & Angstadt, M. (2015). White matter neuroanatomical differences in young children who stutter. *Brain. A Journal of Neurology*, *138*, 694–711. https://doi.org/10.1093/brain/awu400

Civier, O., Bullock, D., Max, L., & Guenther, F. H. (2009). Simulating neural impairments to syllable-level command generation in stuttering. In *World Congress on Fluency Disorders* (Vol. 6, pp. 354–378). Rio de Janeiro, Brazil.

Civier, O., Bullock, D., Max, L., & Guenther, F. H. (2011a). A neural modeling study of stuttering and fluency enhancement by drugs that partially block dopamine action. In *Congress for People Who Stutter* (Vol. 9). Buenos Aires, Argentina.

Civier, O., Bullock, D., Max, L., & Guenther, F. H. (2011b). Dopamine excess may delay selection of syllabic motor programs: A modeling study of stuttering. In *International Congress of Phonetic Sciences (ICPhS)* (Vol. 17, pp. 504–507). Hong Kong, China.

Cleland, J., Lloyd, S., Campbell, L., Crampin, L., Palo, J. P., E, S., … Zharkova, N. (2019). The impact of real-time articulatory information on phonetic transcription: ultrasound-aided transcription in cleft lip and palate speech. *Folia Phoniatrica et Logopaedica*, 1–11.

Cleland, J., Scobbie, J. M., Heyde, C. J., Roxburgh, Z., & Wrench, A. A. (2017). Covert contrast and covert errors in persistent velar fronting. *Clinical Linguistics & Phonetics*, *31*(1), 35–55. https://doi.org/10.1080/02699206.2016.1209788

Cleland, J., Scobbie, J. M., Roxburgh, Z., & Heyde, C. J. (2015). Ultrasound visual biofeedback for heterogeneous persistent speech sound disorders: The UltraPhonix Project. In *Ultrafest VII*.

Cleland, J., Scobbie, J. M., Roxburgh, Z., & Heyde, C. J. (2016). UltraPhonix : learning new articulations with ultrasound.

Connally, E. L., Ward, D., Howell, P., & Watkins, K. E. (2014). Disrupted white matter in language and motor tracts in developmental stuttering. *Brain and Language*, *131*, 25–35. https://doi.org/10.1016/j.bandl.2013.05.013

Conover, W. J., Johnson, M. E., & Johnson, M. M. (1981). A Comparative Study of Tests for Homogeneity of Variances, with Applications to the Outer Continental Shelf Bidding Data. *Technometrics*, *23*(4), 351. https://doi.org/10.2307/1268225

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: hesitations in speech affect language comprehension. *Cognition*, *105*(3), 658–668. https://doi.org/10.1016/j.cognition.2006.10.010

Costa, D., & Kroll, R. (2000). Stuttering: an update for physicians. *Canadian Medical Association Journal*, *162*(13), 1849–1855.

Craig, A., Hancock, K., Tran, Y., Craig, M., & Peters, K. (2002). Epidemiology of stuttering in the community across the entire life span. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1097–1105.

Cross, D. E., & Luper, H. L. (1979). Voice Reaction Time of stuttering and nonstuttering children and adults. *Journal of Fluency Disorders*, *4*, 59–77.

Crystal, T. H., & House, A. S. (1990). Articulation rate and the duration of syllables and stress groups in connected speech. *The Journal of the Acoustical Society of America*, *88*(1), 101–112.

Daniloff, R. G., & Hammarberg, R. (1973). On defining coarticulation. *Journal of Phonetics*, *1*(3), 239–248.

Daniloff, R. G., & Moll, K. L. (1968). Coarticulation of lip rounding. *Journal of Speech, Language, and Hearing Research*, *11*(4), 707–721.

Davidson, L. (2005). Addressing phonological questions with ultrasound, *19*(September 2004), 619–633. https://doi.org/10.1080/02699200500114077

Davidson, L. (2007). Coarticulation in contrastive Russian stop sequences. In *International Congress of Phonetic Sciences (ICPhS)* (Vol. 16, pp. 417–420).

Davis, B. L., & MacNeilage, P. F. (1995). The articulatory basis of babbling. *Journal of Speech, Language, and Hearing Research*, *38*(6), 1199. https://doi.org/10.1044/jshr.3806.1199

Dayalu, V. N., Saltuklaroglu, T., Kalinowski, J., Stuart, A., & Rastatter, M. P. (2001). Producing the vowel /a/ prior to speaking inhibits stuttering in adults in the English language. *Neuroscience Letters*, *306*(1–2), 111–115.

De Nil, L. F., & Brutten, G. J. (1991). Voice onset times of stuttering and nonstuttering children: The influence of externally and linguistically imposed time pressure. *Journal of Fluency Disorders*, *16*(2), 143–158.

Dehqan, A., Yadegari, F., Blomgren, M., & Scherer, R. C. (2016). Formant transitions in the fluent speech of Farsi-speaking people who stutter. *Journal of Fluency Disorders*, *48*, 1–15. https://doi.org/10.1016/J.JFLUDIS.2016.01.005

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321. https://doi.org/10.1037/0033-295X.93.3.283

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science*, *23*(4), 517–542. https://doi.org/10.1207/s15516709cog2304_6

Derrick, D., Best, C. T., & Fiasson, Ro. (2015). *Non-metallic ultrasound probe holder for co-collection and co-registration with EMA*.

Di Simoni, F. G. (1974). Preliminary study of certain timing relationships in the speech of stutterers. *The Journal of the Acoustical Society of America*, *56*(2), 695–696. https://doi.org/10.1121/1.1903313

Dokoza, K. P., Hedever, M., & Sarić, J. P. (2011). Duration and variability of speech segments in fluent speech of children with and without stuttering. *Collegium Antropologicum*, *35*(2), 281–288.

Durand, J. (2004). English in early 21st century Scotland: A phonological perspective. *Tribune International e Des Langues Vivantes*, *36*, 87–105.

Dworzynski, K., Howell, P., Au-Yeung, J., & Rommel, D. (2004). Stuttering on function and content words across age groups of German speakers who stutter. *Journal of Multilingual Communication Disorders*, *2*(2), 81–101.

Falk, S., Maslow, E., Thum, G., & Hoole, P. (2016). Temporal variability in sung productions of adolescents who stutter. *Journal of Communication Disorders*, *62*, 101–114. https://doi.org/10.1016/J.JCOMDIS.2016.05.012

Fant, G. C. M. (1966). A note on vocal tract size factors and non-uniform F-pattern scalings. *Speech Transmission Laboratory Quarterly Progress and Status Report*, *1*, 22–30.

Fant, G. C. M. (1970). *Acoustic theory of speech production: With calculations based on X-ray studies of Russian articulations* (Vol. 2). The Hague: Mouton & Co N. V.

Farnetani, E., & Recasens, D. (1999). Coarticulation models in recent speech production theories. In *Coarticulation: Theory, Data and Techniques* (pp. 31–65).

Feng, Y., & Max, L. (2014). Accuracy and precision of a custom camera-based system for 2-d and 3-d motion tracking during speech and nonspeech motor tasks. *Journal of Speech, Language, and Hearing Research : JSLHR*, *57*(2), 426–438. https://doi.org/10.1044/2014_JSLHR-S-13-0007

Fitch, W. . T. ., & Giedd, J. . (1999). Morphology and development of the human vocal tract : A study using magnetic resonance imaging. *Journal of Acoustical Society of America*, *106*(October), 1511–1522. https://doi.org/10.1121/1.427148

Fowler, C. A. (1980). Coarticulation and theories of extrinsic timing. *Journal of Phonetics*, *8*(1), 113–133.

Fowler, C. A. (2000). Coarticulation resistance of American English consonants and its effects on transconsonantal vowel to vowel coarticulation. *Language and Speech*, *43*(1), 1–41.

Fowler, C. A., Rubin, P., Remez, R. E., & Turvey, M. E. (1980). Implications for speech production of a general theory of action. *Language in Production, Speech and Talk*, *1*, 373–420.

Fowler, C. A., & Saltzman, E. L. (1993). Coordination and coarticulation in speech production. *Language and Speech*, *36*(2, 3), 171–195. https://doi.org/10.1177/002383099303600304

Fraser, M. (2007). *Self-therapy for the Stutterer* (10th ed.). The Stuttering Foundation of America.

Frisch, S. A. (2010). Semi-automatic measurement of stop-consonant articulation using edgetrak. In *Ultrafest* (Vol. 5).

Frisch, S. A., & Maxfield, N. D. (2017). Velar-vowel coarticulation across the lifespan and in people who stutter: Findings and model. *The Journal of the Acoustical Society of America*, *141*(5), 3747–3747. https://doi.org/10.1121/1.4988258

Frisch, S. A., Maxfield, N. D., & Belmont, A. J. (2016). Anticipatory coarticulation and stability of speech in typically fluent speakers and people who stutter. *Clinical Linguistics & Phonetics*, *30*(3–5), 277–291. https://doi.org/10.3109/02699206.2015.1137632

Frisch, S. A., & Wodzinski, S. M. (2016). Velar-vowel coarticulation in a virtual target model of stop production. *Journal of Phonetics*, *56*, 52–65. https://doi.org/10.1016/j.wocn.2016.01.001

Fromkin, V. A. (1971). The Non-Anomalous Nature of Anomalous Utterances, *47*(1), 27–52.

Fuchs, S., Perrier, P., Geng, C., & Mooshammer, C. (2006). What role does the palate play in speech motor control? Insights from tongue kinematics for German alveolar obstruents. *Speech Production: Models, Phonetic Processes, and Techniques*, 149–164.

Fuchs, S., Petrone, C., Krivokapić, J., & Hoole, P. (2013). Acoustic and respiratory evidence for utterance planning in German. *Journal of Phonetics*, *41*(1), 29–47. https://doi.org/10.1016/j.wocn.2012.08.007

Garrett, M. F. (1975). The Analysis of Sentence Production. *Psychology of Learning and Motivation*, *9*, 133–177. https://doi.org/10.1016/S0079-7421(08)60270-4

Geitz, B. E. (1998). *The development of stop consonant place of articulation in preadolescent children*. Vanderbilt University.

Geng, C. C., Turk, A., Scobbie, J. M., Macmartin, C., Hoole, P., Richmond, K., … Campbell, Z. (2013). Recording speech articulation in dialogue: Evaluating a synchronized double Electromagnetic Articulography setup. *Journal of Phonetics*, *41*(6), 421–431. https://doi.org/10.1016/j.wocn.2013.07.002

Gibbon, F. E., Hardcastle, W. J., Crampin, L., Reynolds, B., Razzell, R., & Wilson, J. (2001). Visual feedback therapy using electropalatography (EPG) for articulation disorders associated with cleft palate. *Asia Pacific Journal of Speech, Language and Hearing*, *6*(1), 53–58. https://doi.org/10.1179/136132801805576798

Gibbon, F. E., & Wood, S. E. (2010). Visual feedback therapy with electropalatography. In A. L. Williams, S. McLeod, & R. J. McCauley (Eds.), *Interventions for Speech Sound Disorders in Children* (pp. 509–536). Baltimore: Paul H. Brookes Pub.

Gick, B. (2002). An X-ray investigation of pharyngeal constriction in American English schwa. *Phonetica*, *59*(1), 38–48. https://doi.org/10.1159/000056204

Gick, B., Allen, B., Roewer-Després, F., & Stavness, I. (2017). Speaking tongues are actively braced. *Journal of Speech Language and Hearing Research*, *60*(3), 494. https://doi.org/10.1044/2016_JSLHR-S-15-0141

Gick, B., Allen, B., Stavness, I., & Wilson, I. (2013). Speaking tongues are always braced. In *Journal of Acoustical Society of America* (Vol. 134, p. 10). https://doi.org/10.1002/cnm.2486.Fels

Gick, B., & Campbell, F. (2003). Intergestural timing in English /r/. In *International Congress of Phonetic Sciences (ICPhS)* (Vol. 15, pp. 1–4).

Gick, B., Wilson, I., & Derrick, D. (2012). *Articulatory phonetics*. Malden, MA: John Wiley & Sons Inc.

Giegerich, H. (1992). *English phonology: An introduction*. Cambridge: Cambridge University Press.

Goldstein, L., Chitoran, I., & Selkirk, E. (2007). Syllable structure as coupled oscillator modes: Evidence from Georgian vs. Tashlhiyt Berber. In *Congress of Phonetic Sciences* (Vol. 16, pp. 241–244).

Goldstein, L., Nam, H., Saltzman, E. L., & Chitoran, I. (2009). Coupled oscillator planning model of speech timing and syllable structure. *Frontiers in Phonetics and Speech Sciences*, 239–250.

Goldstein, L., Pouplier, M., Chen, L., Saltzman, E. L., & Byrd, D. (2007). Dynamic action units slip in speech production errors. *Cognition*, *103*(3), 386–412. https://doi.org/10.1016/j.cognition.2006.05.010

Gracco, V. L., & Abbs, J. H. (1986). Variant and invariant characteristics of speech movements. *Experimental Brain Research*, *65*(1), 156–166. https://doi.org/10.1007/BF00243838

Grice, M., Hermes, A., Lancia, L., Mücke, D., Zharkova, N., Lickley, R. J., & Hardcastle, W. J. (2014). Development of lingual coarticulation and articulatory constraints between childhood and adolescence: An ultrasound study. In *ISSP* (pp. 472–475).

Guenther, F. H. (1994). A Neural Network Model Of Speech Acquisition And Motor Equivalent Speech Production Running title: Speech acquisition and motor equivalence. *Biological Cybernetics*, *72*, 43–53. https://doi.org/10.1007/BF00206237

Guenther, F. H., & Guenther, H. (1995). Speech Sound Acquisition, Coarticulation, and Rate Effects in a Neural Network Model of Speech Production. *Psychological Review*, *102*(3), 594–621.

Guitar, B. E. (2013). *Stuttering: An Integrated Approach to Its Nature and Treatment*. Baltimore: Lippincott Williams & Wilkins.

Guitar, B. E., & Belin-Frost, G. (1998). *Stuttering. An integrated approach to its nature and treatment*. Baltimore: Williams & Wilkins.

Haken, H., Peper, C. E., Beek, P. J., & Daffertshofer, A. (1996). A model for phase transitions in human hand movements during multifrequency tapping. *Physica D: Nonlinear Phenomena*, *90*(1–2), 179–196.

Harbison, D. C., Porter, R. J., & Tobey, E. A. (1989). Shadowed and simple reaction times in stutterers and nonstutterers. *The Journal of the Acoustical Society of America*, *86*(4), 1277–1284. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2808903

Healey, E. C., Mallard, A. R., & Adams, M. R. (1976). Factors Contributing to the Reduction of Stuttering during Singing. *Journal of Speech and Hearing Research*, *19*(3), 475–480. https://doi.org/10.1044/jshr.1903.475

Healey, E. C., & Ramig, P. R. (1986). Acoustic measures of stutterers' and nonstutterers' fluency in two speech contexts. *Journal of Speech and Hearing Research*, *29*(3), 325–331.

Henke, W. L. (1966). *Dynamic articulatory model of speech production using computer*. Cambridge, MA, US.

Heyde, C. J., Cleland, J., Scobbie, J. M., & Roxburgh, Z. (2016). UltraPhonix: Das Erlernen von Artikulatorischen Gesten mit Ultraschall-Biofeedback. In *Poster presented at 10. Herbsttreffen Patholinguistik*. Potsdam.

Heyde, C. J., Cleland, J., Scobbie, J. M., & Roxburgh, Z. (2017). UltraPhonix: Das Erlernen von Artikulatorischen Gesten mit Ultraschall-Biofeedback. *Spektrum Patholinguistik*, *10*, 197–208.

Heyde, C. J., & Scobbie, J. M. (2016). Wenn Stotterer nicht stottern. Quantifizierung dynamischer Ultraschalldaten. In *P und P 12* (pp. 63–65). Munich. https://doi.org/10.1109/ICCSE.2014.6926416

Heyde, C. J., Scobbie, J. M., & Finlayson, I. R. (2015). Searching for Closure: Seeing a Dip. In *Poster presented at Ultrafest VII*. Hong Kong.

Heyde, C. J., Scobbie, J. M., Lickley, R., & Drake, E. K. E. (2016). How fluent is the fluent speech of people who stutter ? A new approach to measuring kinematics with ultrasound approach to measuring kinematics with ultrasound. *Clinical Linguistics & Phonetics*, *30*(3–5), 292–312. https://doi.org/10.3109/02699206.2015.1100684

Heyne, M., & Derrick, D. (2015). Benefits of using polar coordinates for working with ultrasound midsagittal tongue contours. In *Poster presented at the 169th meeting of the Acoustical Society of America*. Pittsburgh, PA.

Hoole, Phil, & Harrington, J. (2013). Measuring physiological speech entrainment with electromagnetic articulography (EMA).

Hoole, Philip, & Nguyen, N. (1997). Electromagnetic articulography in coarticulation research. *Forschungsberichte Des Instituts Für Phonetik Und Sprachliche Kommunikation Der Universität München*, *35*, 177–184.

Horii, Y. (1984). Phonatory initiation, termination, and vocal frequency change reaction times of stutterers. *Journal of Fluency Disorders*, *9*(2), 115–124. https://doi.org/10.1016/0094-730X(84)90029-9

Hourcade, J. P., Bederson, B. B., Druin, A., & Guimbretière, F. (2004). Differences in pointing task performance between preschool children and adults using mice. *ACM Transactions on Computer-Human Interaction*, *11*(4), 357–386. https://doi.org/10.1145/1035575.1035577

Howard, S., & Varley, R. (1995). EPG in Therapy Using electropalatography to treat severe acquired apraxia of speech. *International Journal of Language & Communication Disorders*, *30*(2), 246–255. https://doi.org/10.3109/13682829509082535

Howell, P. (2004). Assessment of some contemporary theories of stuttering that apply to spontaneous speech. *Contemporary Issues in Communication Science and Disorders: CICSD*, *31*, 122–139.

Howell, P. (2007). Signs of developmental stuttering up to age eight and at 12 plus. *Clinical Psychology Review*, *27*(3), 287–306. https://doi.org/10.1016/j.cpr.2006.08.005

Howell, P., & Au-Yeung, J. (2000). The EXPLAN theory of fluency control applied to the diagnosis of stuttering. *Amsterdam Studies in the Theory and History of Linguistic Science Series 4*, (1989), 75–94.

Howell, P., Au-Yeung, J., & Sackin, S. (1999). Exchange of Stuttering From Function Words to Content Words With Age. *Journal of Speech, Language, and Hearing Research*, *42*(2), 345–354. https://doi.org/10.1044/jslhr.4202.345

Howell, P., Ruffle, L., Fernandez-Zuniga, A., Gutierrez, R., Fernandez A H, O'Brian, M. L., … Au-Yeung, J. (2004). Comparison of exchange patterns of stuttering in Spanish and English monolingual speakers and a bilingual Spanish-English speaker. *Theory, Research and Therapy in Fluency Disorders*, *415–422*.

Howell, P., & Williams, M. (1992). Acoustic analysis and perception of vowels in children's and teenagers' stuttered speech. *The Journal of the Acoustical Society of America*, *91*(3), 1697–1706. https://doi.org/10.1121/1.402449

Howell, P., Williams, M., & Vause, L. (1987). Acoustic analysis of repetitions in stutterers' speech. In *Speech Motor Dynamics in Stuttering* (pp. 371–380). Vienna: Springer Vienna. https://doi.org/10.1007/978-3-7091-6969-8_29

Howie, P. M. (1981). Concordance for stuttering in monozygotic and dizygotic twin pairs. *Journal of Speech, Language, and Hearing Research*, *24*(3), 317–321.

Hubbard, C. P., & Prins, D. (1994). Word familiarity, syllabic stress pattern, and stuttering. *Journal of Speech, Language, and Hearing Research*, *37*(3), 564–571.

Hubbard, C. P., & Yairi, E. (1988). Clustering of disfluencies in the speech of stuttering and nonstuttering preschool children. *Journal of Speech, Language, and Hearing Research*, *31*(2), 228–233.

International Expert Panel on Multilingual Children's Speech. (2012). *Multilingual children with speech sound disorders: Position paper*. Bathurst, Australia.

Iskarous, K. (2005a). Detecting the edge of the tongue: A tutorial. *Clinical Linguistics & Phonetics*, *19*(6), 555–565. https://doi.org/10.1080/02699200500113871

Iskarous, K. (2005b). Patterns of tongue movement. *Journal of Phonetics*, *33*(4), 363–381. https://doi.org/10.1016/j.wocn.2004.09.001

Iskarous, K. (2013). The development of motor synergies in children : Ultrasound and acoustic measurements, *133*(1).

Iskarous, K., Fowler, C. A., & Whalen, D. H. (2010). Locus equations are an acoustic expression of articulator synergy. *Acoustical Society of America*, *128*(4), 2021–2032. https://doi.org/10.1121/1.3479538

Iverach, L., & Rapee, R. M. (2014). Social anxiety disorder and stuttering: Current status and future directions. *Journal of Fluency Disorders*. https://doi.org/10.1016/j.jfludis.2013.08.003

Jacks, A., & Haley, K. L. (2015). Auditory masking effects on speech fluency in apraxia of speech and aphasia: Comparison to altered auditory feedback. *Journal of Speech, Language, and Hearing Research*, *58*(6), 1670–1686.

Jackson, E. S., Tiede, M., Beal, D., & Whalen, D. H. (2016). The Impact of Social–Cognitive Stress on Speech Variability, Determinism, and Stability in Adults Who Do and Do Not Stutter. *Journal of Speech, Language, and Hearing Research*, *59*(6), 1295–1314. https://doi.org/10.1044/2016_JSLHR-S-16-0145

Jackson, E. S., Yaruss, J. S., Quesal, R. W., Terranova, V., & Whalen, D. H. (2015). Responses of adults who stutter to the anticipation of stuttering. *Journal of Fluency Disorders*, *45*, 38–51. https://doi.org/10.1016/j.jfludis.2015.05.002

Jäncke, L. (1994). Variability and duration of voice onset time and phonation in stuttering and nonstuttering adults. *Journal of Fluency Disorders*, *19*(1), 21–37. https://doi.org/10.1016/0094-730X(94)90012-4

Jansen-Osmann, P., Richter, S., Konczak, J., & Kalveram, K.-T. (2002). Force adaptation transfers to untrained workspace regions in children. *Experimental Brain Research*, *143*(2), 212–220. https://doi.org/10.1007/s00221-001-0982-8

Johnson, W. (1930). *Because I stutter*. New York, London: D. Appelton and Company.

Johnson, W. (1959). *The onset of stuttering research findings and implications*. Oxford, England: University of Minnesota Press.

Johnson, W. (1961). Measurements of oral reading and speaking rate and disfluency of adult male and female stutterers and nonstutterers. *Journal of Speech and Hearing Disorders. Monograph Supplement*, *7*, 1–20.

Kelso, J. A., Saltzman, E. L., & Tuller, B. (1986). The dynamical perspective on speech production: Data and theory. *Journal of Phonetics*, *14*(1), 29–59.

Kempen, G. (1987). An Incremental Proceduml Grammar for Sentence Formulation, *258*, 201–258.

Kent, R. D. (1983). The segmental organization of speech. In *The Production of Speech* (pp. 57–89). New York, NY: Springer New York. https://doi.org/10.1007/978-1-4613-8202-7_4

Kent, R. D. (2015). Nonspeech oral movements and oral motor disorders: A narrative review. *American Journal of Speech-Language Pathology*, *24*(4), 763–789. https://doi.org/10.1044/2015

Kent, R. D., & Minifie, F. D. Coarticulation in recent speech production models, 5 Journal of Phonetics § (1977). https://doi.org/10.1111/j.1460-9568.2009.06973.x

Kent, R. D., & Moll, K. L. (1969). Vocal-Tract Characteristics of the Stop Cognates. *The Journal of the Acoustical Society of America*, *46*(6B), 1549–1555. https://doi.org/10.1121/1.1911902

Koopmans-Van Beinum, F. J. (1993). Cyclic effects of infant speech perception, early sound production, and maternal speech. In *Proceedings of the Institute of Phonetic Sciences* (pp. 65–78). Amsterdam.

Kozhevnikov, V. A., & Christovich, L. A. (1965). Speech: Articulation and perception. *Joint Publications Research Service*, *30*.

Krakow, R. A. (1999). Physiological organization of syllables: A review. *Journal of Phonetics*, *27*(1), 23–54. https://doi.org/10.1006/JPHO.1999.0089

Kröger, B. J. (2013). Modeling of speech production from the perspective of neuroscience. *Systemtheorie, Signalverarbeitung Und Sprachtechnologie*, 218–225.

Kroos, C., Hoole, P., Kühnert, B., & Tillmann, H. G. (1997). Phonetic evidence for the phonological status of the tense-lax distinction in German. *Journal of the Acoustical Society of America*, *100*(35), 17–25.

Krull, D. (1987). Second formant locus patterns as a measure of consonant-vowel coarticulation. In *Phonetic Experimental Research at the Institute of Linguistics at the University of Stockholm (PERILUS)* (Vol. 5, pp. 43–61). Stockholm: University of Stockholm.

Krull, D. (1988). Acoustic properties as predictors of perceptual responses: A study of Swedish voiced stops. In *Phonetic Experimental Research at the Institute of Linguistics at the University of Stockholm (PERILUS)* (pp. 66–70). Stockholm: University of Stockholm.

Krull, D. (1989a). *Consonant-vowel coarticulation in spontaneous speech and in reference words*. *Speech Transmission Laboratory, Quaterly status and progress report 2/1989*. Stockholm.

Krull, D. (1989b). Second formant locus patterns and consonant-vowel coarticulation in spontaneous speech. In O. Engstrand, M. Dufberg, & C. Kylander (Eds.), *Phonetic Experimental Research at the Institute of Linguistics at the University of Stockholm (PERILUS)* (Vol. 10, pp. 87–108). Stockholm: University of Stockholm.

Kühnert, B., Hoole, P., & Mooshammer, C. (2006). Gestural overlap and C-center in selected French consonant clusters. In *Proceedings of the 7th Speech Production Seminar* (Vol. 7, pp. 40–48). Ubatuba, Brazil. Retrieved from papers2://publication/uuid/584AC5A6-95F9-474E-A736-4265BE57A1B0

Kühnert, B., & Nolan, F. (1999). The origin of coarticulation. *Coarticulation Theory Data And*, 7–30. https://doi.org/10.1017/CBO9780511486395.002

Ladefoged, P. (2006). *A Course in Phonetics* (5th ed.). Thomson, Wadsworth.

Lambert, J., & Bard, C. (2005). Acquisition of visuomanual skills and improvement of information processing capacities in 6- to 10-year-old children performing a 2D pointing task. *Neuroscience Letters*, *377*(1), 1–6. https://doi.org/10.1016/J.NEULET.2004.11.058

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2013). Bunched /r/ promotes vowel merger to schwar: An ultrasound tongue imaging study of Scottish sociophonetic variation. *Journal of Phonetics*, *41*(3–4), 198–210. https://doi.org/10.1016/J.WOCN.2013.01.004

Lawson, E., Scobbie, J. M., & Stuart-Smith, J. (2014). A socio-articulatory study of Scottish rhoticity. In *Sociolinguistics in Scotland* (pp. 53–78).

Lawson, E., Stuart-Smith, J., & Scobbie, J. M. (2014). A mimicry study of adaptation towards socially-salient tongue shape variants. *University of Pennsylvania Working Papers in Linguistics*, *20*(2), 99–110.

Lawson, E., Stuart-Smith, J., Scobbie, J. M., & Nakai, S. (2015). *Seeing Speech: An articulatory web resource for the study of phonetics*. University of Glasgow.

Lehiste, I., & Peterson, G. E. (1959). Vowel Amplitude and Phonemic Stress in American English. *The Journal of the Acoustical Society of America*, *31*(4), 428–435. https://doi.org/10.1121/1.1907729

Leisman, G., Zenhausern, R., Ferentz, A., Tefera, T., & Zemcov, A. (1995). Electromyographic Effects of Fatigue and Task Repetition on the Validity of Estimates of Strong and Weak Muscles in Applied Kinesiological Muscle-Testing Procedures. *Perceptual and Motor Skills*, *80*(3), 963–977. https://doi.org/10.2466/pms.1995.80.3.963

Lenoci, G. (2018). Anticipatory coarticulation in the speech of people who stutter. https://doi.org/10.17 4 69 / O2103AISV000020

Lenoci, G., & Ricci, I. (2017). An ultrasound study of anticipatory coarticulation in the speech of Italian children who stutter. *International Symposium of Monolingual and Bilingual Speech*.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition*, *14*(1), 41–104.

Levelt, W. J. M. (1984). Spontaneous self-repairs in speech: Processes and representations. In *Proceedings of the tenth international congress of phonetic science* (pp. 105–117).

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levelt, W. J. M. (1992). Accessing words in speech production: processes, stages and representations. *Cognition*, *42*(1–3), 1–22.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Li, M., Kambhamettu, C., & Stone, M. (2005). Automatic contour tracking in ultrasound images. *Clinical Linguistics & Phonetics*, *19*(6–7), 545–554. https://doi.org/10.1080/02699200500113616

Liberman, A. M. (1970). Some characteristics of perception in the speech mode. *Perception and Its Disorders*, *18*, 238–254.

Lickley, R. (2017). Disfluency in typical and stuttered speech. In *Fattori sociali e biologici nella variazione fonetica - Social and biological factors in speech variation* (3rd ed., pp. 373–387). Milano. https://doi.org/10.17469/O2103AISV000019

Lickley, R. J. (2015). Fluency and Disfluency. In *The Handbook of Speech Production* (pp. 445–474).

Lickley, R. J., Hartsuiker, R. J., Corley, M., Russell, M., & Nelson, R. (2005). Judgment of disfluency in people who stutter and people who do not stutter: Results from magnitude estimation. *Language and Speech*, *48*(3), 299–312.

Lin, Y., & Mielke, J. (2008). Discovering Place and Manner Features: What Can Be Learned from Acoustic and Articulatory Data? *University of Pennsylvania Working Papers in Linguistics*, *14*(1).

Lindblom, B. (1963). Spectrographic study of vowel reduction. *Journal of Acoustical Society of America*, *35*(11), 1773/1781.

Lindblom, B. (1983). Economy of speech gestures. In P. F. MacNeilage (Ed.), *The Production of Speech* (pp. 217–245). New York: Springer New York. https://doi.org/10.1007/978-1-4613-8202-7_10

Lindblom, B. (1990). The status of phonetic gestures. In *Perilus XI* (pp. 21–40). Stockholm.

Lindblom, B., & Sussman, H. M. (2012). Dissecting coarticulation: How locus equations happen. *Journal of Phonetics*, *40*(1), 1–19. https://doi.org/10.1016/j.wocn.2011.09.005

Liss, J. M., & Weismer, G. (1992). Qualitative acoustic analysis in the study of motor speech disorders. *The Journal of the Acoustical Society of America*, *92*(5), 2984–2987. https://doi.org/10.1121/1.404364

Loucks, T. M. J., & De Nil, L. F. (2012). Oral sensorimotor integration in adults who stutter. *Folia Phoniatrica et Logopaedica : Official Organ of the International Association of Logopedics and Phoniatrics (IALP)*, *64*(3), 116–121. https://doi.org/10.1159/000338248 [doi]

Luce, P. A., & Charles-Luce, J. (1985). Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. *The Journal of the Acoustical Society of America*, *78*(6), 1949–1957. https://doi.org/10.1121/1.392651

Ludlow, C. L., & Loucks, T. M. J. (2003). Stuttering: a dynamic motor control disorder. *Journal of Fluency Disorders*, *28*(4), 273–295. https://doi.org/10.1016/j.jfludis.2003.07.001

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19–44.

MacNeilage, P. F., & DeClerk, J. L. On the Motor Control of Coarticulation in CVC Monosyllables, 45 The Journal of the Acoustical Society of America § (1969). Acoustical Society of America. https://doi.org/10.1121/1.1911593

MacPherson, M. K., & Smith, A. (2013). Influences of sentence length and syntactic complexity on the speech motor control of children who stutter. *Journal of Speech Language and Hearing Research*, *56*(1), 89. https://doi.org/10.1044/1092-4388(2012/11-0184)

Manning, W. H., & DiLollo, A. (2017). *Clinical decision making in fluency disorders* (Fourth Edi). Plural Publishing.

Månsson, H. (2000). Childhood stuttering: Incidence and development. *Journal of Fluency Disorders*, *25*(1), 47–57.

Maruthy, S., Feng, Y., & Max, L. (2017). Spectral Coefficient Analyses of Word-Initial Stop Consonant Productions Suggest Similar Anticipatory Coarticulation for Stuttering and Nonstuttering Adults. *Language and Speech*, *61*(1), 31–42. https://doi.org/10.1177/0023830917695853

Max, L., & Gracco, V. L. (2005). Coordination of oral and laryngeal movements in the perceptually fluent speech of adults who stutter. *Journal of Speech, Language,*

and Hearing Research, *48*(3), 524–542.

Max, L., Guenther, F. H., Gracco, V. L., Ghosh, S. S., & Wallace, M. E. (2004). Unstable or insufficiently activated internal models and feedback-biased motor control as sources of dysfluency: A theoretical model of stuttering. *Contemporary Issues in Communication Science and Disorders*, *31*, 105–122.

Mayo, C., & Turk, A. (2004). Adult-child differences in acoustic cue weighting are influenced by segmental context: children are not always perceptually biased toward transitions. *The Journal of the Acoustical Society of America*, *115*(6), 3184–3194.

McCann, J., Timmins, C., Wood, S. E., Hardcastle, W. J., & Wishart, J. G. (2009). Electropalatographic therapy for children and young people with Down's syndrome. *Clinical Linguistics & Phonetics*, *23*(12), 926–939.

McClean, M. D., & Levandowski, D. R. (1994). Intersyllabic movement timing in the fluent speech of stutters with different disfluency levels. *Journal of Speech & Hearing Research*, *37*(5).

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. https://doi.org/10.1037/0033-295X.88.5.375

Melnick, K. S., Conture, E. G., & Ohde, R. N. (2003). Phonological priming in picture naming of young children who stutter. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1428. https://doi.org/10.1044/1092-4388(2003/111)

Meyer, D. E., & Gordon, P. C. (1985). Speech production: Motor programming of phonetic features. *Journal of Memory and Language*, *24*(1), 3–26. https://doi.org/10.1016/0749-596X(85)90013-0

Moll, K. L., & Daniloff, R. G. (1971). Investigation of the timing of velar movements during speech. *The Journal of the Acoustical Society of America*, *50*(2b), 678–684.

Montgomery, A. A., Reed, P. E., Crass, K. A., Hubbard, H. I., & Stith, J. (2014). The effects of measurement error and vowel selection on the locus equation measure of coarticulation. *Acoustical Society of America*, *136*(5), 2747–2750. https://doi.org/10.1121/1.4896460

Mooshammer, C., Goldstein, L., Nam, H., McClure, S., Saltzman, E. L., & Tiede, M. (2012). Bridging planning and execution : Temporal planning of syllables. *Journal of Phonetics*, *40*(3), 374–389. https://doi.org/10.1016/j.wocn.2012.02.002

Mooshammer, C., Hoole, P., & Kühnert, B. (1995). On loops. *Journal of Phonetics*, *23*(1–2), 3–21. https://doi.org/10.1016/S0095-4470(95)80029-8

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6*(4), 413–429. https://doi.org/10.1111/1467-7687.00296

Munhall, K. G., & Vatikiotis-Bateson, E. (1998). The moving face during speech communication. In *Hearing by eye II: Advances in the psychology of speechreading and auditory-visual speech* (pp. 123–139).

Murdoch, B. E., Theodoros, D. G., Stokes, P. D., & Goozée, J. V. (2000). Kinematic analysis of tongue movements in dysarthria following traumatic brain injury using electromagnetic articulography. *Brain Injury*, *14*(2), 153–174.

Natke, U., Gosser, J., Sandriseser, P., & Kalveram, K.-T. (2002). The duration component of the stress effect in stuttering. *Journal of Fluency Disorders*, *27*(4), 305–318.

Nearey, T. M., & Shammass, S. E. (1987). Formant transitions as partly distinctive invariant properties in the identification of voiced stops. *Canadian Acoustics*, *15*(4), 17–24.

Neef, N. E., Bütfering, C., Anwander, A., Friederici, A. D., Paulus, W., & Sommer, M. (2016). Left posterior-dorsal area 44 couples with parietal areas to promote speech fluency, while right area 44 activity promotes the stopping of motor responses. *NeuroImage*, *142*, 628–644. https://doi.org/10.1016/j.neuroimage.2016.08.030

Newman, P. (1972). Syllable weight as a phonological variable. *Studies in African Linguistics*, *3*(3), 301–318.

Nippold, M. A. (2001). Phonological disorders and stuttering in children: What is the frequency of co-occurrence? *Clinical Linguistics & Phonetics*, *15*(3), 219–228. https://doi.org/10.1080/02699200010010523

Nittrouer, S. (2006). Children hear the forest. *The Journal of the Acoustical Society of America*, *120*(4), 1799–1802. https://doi.org/10.1121/1.2335273

Nittrouer, S., Studdert-Kennedy, M., & Mcgowan, R. S. (1989). The emergence of phonetic segments: Evidence from the spectral structure of fricative-vowel syllables spoken by children and adults. *Journal of Speech, Language, and Hearing Research*, *32*(1), 120–132.

Norrick, N. R. (2009). Interjections as pragmatic markers. *Journal of Pragmatics*, *41*(5), 866–891. https://doi.org/10.1016/J.PRAGMA.2008.08.005

Nudelman, H. B., Herbrich, K. E., Hoyt, B. D., & Rosenfield, D. B. (1987). Dynamic Characteristics of Vocal Frequency Tracking in Stutterers and Nonstutterers. In *Speech motor dynamics in stuttering* (pp. 161–169). Wien: Springer.

Öhman, S. E. G. (1966). Coarticulation in VCV utterances: Spectrographic measurements. *The Journal of the Acoustical Society of America*, *39*(1), 151–168.

Olander, L., Smith, A., & Zelaznik, H. N. (2010). Evidence that a motor timing deficit is a factor in the development of stuttering. *Journal of Speech, Language, and Hearing Research*, *53*(4), 876–886. https://doi.org/10.1044/1092-4388(2009/09-0007)

Öller Darelid, M., Hartelius, L., & Lohmander, A. (2016). Generalised EPG treatment effect in a cochlear implant user maintained after 2 years. *International Journal of Speech-Language Pathology*, *18*(1), 65–76. https://doi.org/10.3109/17549507.2015.1048827

Onslow, M., Van Doorn, J., & Newman, D. (1992). Variability of acoustic segment durations after prolonged-speech treatment for stuttering. *Journal of Speech and Hearing Research*, *35*(3), 529–536.

Oppenheim, G. M., Dell, G. S., & Schwartz, M. F. (2010). The dark side of incremental learning: A model of cumulative semantic interference during lexical access in speech production. *Cognition*, *114*(2), 227–252. https://doi.org/10.1016/J.COGNITION.2009.09.007

Packman, A., Onslow, M., & Doorn, J. van. (1994). Prolonged Speech and Modification of Stuttering. *Journal of Speech, Language, and Hearing Research*, *37*(4), 724–737. https://doi.org/10.1044/jshr.3704.724

Packman, A., Onslow, M., Richard, F., & Van Doorn, J. (1996). Syllabic stress and variability: A model of stuttering. *Clinical Linguistics & Phonetics*, *10*(3), 235–263.

Pape, D., Perrier, P., Fuchs, S., & Kandel, S. (2011). Does physical realism of articulatory modeling improve the perception of synthetic speech? In L. Ménard, S. R. Baum, V. L. Gracco, & D. J. Ostry (Eds.), *International Seminar on Speech Production (ISSP)* (Vol. 9, pp. 153–154). Montréal.

Parks, S. (2001). Moving from school to the workplace. Disciplinary innovation, border crossings, and the reshaping of a written genre. *Applied Linguistics*, *22*(4), 405–438.

Pascoe, M., Stackhouse, J., & Wells, B. (2006). *Persisting speech difficulties in children: children's speech and literacy difficulties*. John Wiley & Sons.

Perkell, J. S., Cohen, M. H., Svirsky, A., Matthies, L., Garabieta, I., & Jackson, M. T. T. (1992). Electromagnetic midsagittal articulometer systems for transducing speech articulatory movements. *The Journal of the Acoustical Society of America*, *92*(6), 3078.

Perkell, J. S., & Klatt, D. (2014). *Invariance and variability in speech processes*. Psychology Press.

Perrier, P., Perkell, J. S., Payan, Y., Zandipour, M., Guenther, F. H., & Khalighi, A. (2007). Degrees of freedom of tongue movements in speech may be constrained by biomechanics. *ArXiv Preprint ArXiv:0709.1405*, 2–5.

Peters, A. M. (1976). *Language learning strategies: does the whole equal the sum of the parts? Papers and Reports on Child Language Development* (Vol. 12).

Peters, H. F. M., Hulstijn, W., & van Lieshout, P. H. H. M. (2000). Recent developments in speech motor research into stuttering. *Folia Phoniatrica et Logopaedica*, *52*(1–3), 103–119. https://doi.org/10.1159/000021518

Pike, K. L., & Pike, E. V. (1947). Immediate Constituents of Mazateco Syllables. *International Journal of American Linguistics*, *13*(2), 78–91. https://doi.org/10.1086/463932

Postma, A. (2000). Detection of errors during speech production: a review of speech monitoring models. *Cognition*, *77*(2), 97–132. https://doi.org/10.1016/S0010-0277(00)00090-1

Postma, A., & Kolk, H. (1993). The Covert Repair Hypothesis: Prearticulatory repair processes in normal and stuttered disfluencies. *Journal of Speech, Language, and Hearing Research*, *36*(3), 472–487.

Postma, A., Kolk, H., & Povel, D. J. (1990). Speech planning and execution in stutterers. *Journal of Fluency Disorders*, *5*(1), 49–59. https://doi.org/10.1016/0094-730X(90)90032-N

Pouplier, M. (2007). *Articulatory perspectives on errors*. *MIT Working Papers in Linguistics* (Vol. 53).

Pouplier, M., & Goldstein, L. (2010). Intention in articulation: Articulatory timing in alternating consonant sequences and its implications for models of speech production. *Language and Cognitive Processes*, *25*(5), 616–649. https://doi.org/10.1080/01690960903395380

Pouplier, M., & Waltl, S. (2008). Articulatory Timing of Coproduced Gestures and Its Implications for Models of Speech Production. In *International Seminar on Speech Production (ISSP)* (Vol. 8, pp. 19–22).

Prasad, V. S. N., Kellokumpu, V., & Davis, L. S. (2006). Ballistic Hand Movements. In *International Conference on Artiuclated motion and Deformable Objects* (pp. 153–164). Springer Berlin Heidelberg. https://doi.org/10.1007/11789239_16

Prasse, J. E., & Kikano, G. E. (2008). Stuttering: an overview. *American Family Physician*, *77*(9), 1271–1276.

Prins, D., Hubbard, C. P., & Krause, M. (1991). Syllabic Stress and the Occurrence of Stuttering. *Journal of Speech and Hearing Research*, *34*(5), 1011–1016. https://doi.org/10.1044/jshr.3405.1011

Proctor, A., Duff, M., & Yairi, E. (2002). Early childhood stuttering: African Americans and European Americans. *ASHA Leader*, *4*(15), 102.

Prosek, R. A., Montgomery, A. A., Walden, B. E., & Hawkins, D. B. (1987). Formant Frequencies of Stuttered and Fluent Vowels. *Journal of Speech, Language, and Hearing Research*, *30*(3), 301. https://doi.org/10.1044/jshr.3003.301

Qin, C., & Carreira-Perpiñán. (2007). An empirical investigation of the nonuniqueness in the acoustic-to-articulatory mapping. In *Eighth Annual Conference of the International Speech Communication Association (Interspeech)* (pp. 74–77).

Recasens, D. (1985). Coarticulatory patterns and degrees of coarticulatory resistance in catalan CV sequences. *Language and Speech*, *28*(2), 97–114. https://doi.org/10.1177/002383098502800201

Recasens, D. (2002). An EMA study of VCV coarticulatory direction. *The Journal of the Acoustical Society of America*, *111*(6), 2828–2841. https://doi.org/10.1121/1.1479146

Recasens, D., & Espinosa, A. (2009). An articulatory investigation of lingual coarticulatory resistance and aggressiveness for consonants and vowels in Catalan. *Acoustical Society of America*, *125*(4), 2288–2298. https://doi.org/10.1121/1.3089222

Reetz, H., & Jongman, A. (2011). *Phonetics: Transcription, production, acoustics, and perception* (34th ed., Vol. 34). John Wiley & Sons.

Reilly, S., Onslow, M., Packman, A., Cini, E., Conway, L., Ukoumunne, O. C., & Wake, M. (2013). Natural history of stuttering to 4 years of age: A prospective community-based study. *Pediatrics*, *132*(3), 460–467.

Remijsen, B. (2013). msr&check_formants_batch. *Praat Script*. Retrieved from http://www.lel.ed.ac.uk/~bert/praatscripts.html

Riley, G. D., & Bakker, K. (2009). SSI-4: Stuttering Severity Instrument. Pro-Ed.

Robb, M. P., & Blomgren, M. (1997). Analysis of F2 transitions in the speech of stutterers and nonstutterers. *Journal of Fluency Disorders*, *22*, 1–16.

Rochet-Capellan, A., & Fuchs, S. (2013). The interplay of linguistic structure and breathing in German spontaneous speech. In *14th Annual Conference of the Internatonal Speech Communication Association (Interspeech)* (p. 1228).

Rosenfelder, I., Fruehwald, J., Evanini, K., & Yuan, J. (2011). FAVE (Forced Alignment and Vowel Extraction) Program Suite.

Rosenfield, D. B. (1980). Cerebral dominance and stuttering *. *Journal of Fluency Disorders*, *5*, 171–185.

RStudio Team. (2015). RStudio: Integrated Development for R. Boston, MA. Retrieved from http://www.rstudio.com/

Rubertus, E., Abakarova, D., Tiede, M., Ries, J., & Noiray, A. (2016). Development of coarticulation in German children : Acoustic and articulatory locus equations. In *Ultrafest* (Vol. 7). Hong Kong, China. https://doi.org/10.13140/RG.2.1.5098.8881

Rudy, K., & Yunusova, Y. (2013). The effect of anatomic factors on tongue position variability during consonants. *Journal of Speech, Language, and Hearing Research*, *56*(1), 137–149. https://doi.org/10.1044/1092-4388(2012/11-0218)a

Rumelhart, D. E. (1998). The architecture of mind: A connectionist approach. In P. Thagard (Ed.), *Mind Readings* (pp. 207–238). Cambridge, MA: MIT Press.

Saltzman, E. L. (1986). *Task dynamic coordination of the speech articulators: A preliminary model*. *Reports-Research/Technical* (Vol. 143). Springfield, VA.

Saltzman, E. L., & Kelso, J. A. (1987). Skilled actions: A task-dynamic approach. *Psychological Review*, *94*(1), 84–106.

Saltzman, E. L., & Munhall, K. G. (1989). A dynamical approach to gestural patterning in speech production. *Ecological Psychology*, *1*(4), 333–382.

Saltzman, E. L., Tyrone, M., & Goldstein, L. (2000). Tutorial on task dynamics and articulatory phonology. In *Articulatory Phonology: Major Hypotheses* (Vol. 1, pp. 1–14). Boston, MA: Haskins Laboratories.

Sander, E. K. (1972). When are speech sounds learned? *Journal of Speech and Hearing Disorders*, *37*(1), 55–63. https://doi.org/10.1044/jshd.3701.55

Sasisekaran, J., De Nil, L. F., Smyth, R., & Johnson, C. (2006). Phonological encoding in the silent speech of persons who stutter. *Journal of Fluency Disorders*, *31*(1), 1–21. https://doi.org/10.1016/j.jfludis.2005.11.005

Sawyer, J., Chon, H., & Ambrose, N. G. (2008). Influences of rate, length, and complexity on speech disfluency in a single-speech sample in preschool children who stutter. *Journal of Fluency Disorders*, *33*(3), 220–240. https://doi.org/10.1016/J.JFLUDIS.2008.06.003

Schegloff, E. A., Jefferson, G., & Sacks, H. (1977). The Preference for Self-Correction in the Organization of Repair in Conversation. *Language*, *53*(2), 361. https://doi.org/10.2307/413107

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language*, *31*(1), 26–35.

Scobbie, J. M. (2018). personal communication.

Scobbie, J. M., Lawson, E., Cowen, S., Cleland, J., & Wrench, A. A. (2011). *A common co-ordinate system for mid-sagittal articulatory measurement*. *QMU CASL Working Papers* (Vol. WP-20).

Scobbie, J. M., Punnoose, R., & Khattab, G. (2013). Articulating five liquids: A single speaker ultrasound study of Malayalam. *Rhotics: New Data and Perspectives*, 99–124.

Scobbie, J. M., Wrench, A. A., & Van Der Linden, M. L. (2008). Head-Probe Stabilisation in Ultrasound Tongue Imaging Using a Headset to Permit Natural Head Movement. In *8th International Seminar on Speech Production (ISSP)* (pp. 373–376).

Sereno, J. A., Baum, S. R., Marean, G. C., & Lieberman, P. (1987). Acoustic analyses and perceptual data on anticipatory labial coarticulation in adults and children. *The Journal of the Acoustical Society of America*, *81*(2), 512–519.

Shaker, H. (2009). *Clinical Physiology of Swallowing Mechanisms*. *Medical Speech and Swallowing Disorders*.

Shapiro, A. I., & Decicco, B. A. (1982). The relationship between normal dysfluency and stuttering: An old question revisited. *Journal of Fluency Disorders*, *7*(1), 109–121. https://doi.org/10.1016/S0094-730X(82)80004-1

Shapiro, D. A. (1999). *Stuttering intervention: A collaborative journey to fluency freedom*. Austin, TX: Pro-Ed.

Shaw, J. A., & Hoole, P. (2011). Dynamic invariance in the phonetic expression of syllable structure : a case study of Moroccan Arabic consonant clusters *, *28*, 455–490. https://doi.org/10.1017/S0952675711000224

Shawker, T. H., Sonies, B. C., & Stone, M. (1984). Soft Tissue Anatomy of the Tongue and Floor of the Mouth : An Ultrasound Demonstration. *National Institutes of Health*, *21*, 335–350.

Sheehan, J. G., & Voas, R. B. (1954). Tension patterns during stuttering in relation to conflict, anxiety-binding, and reinforcement. *Communications Monographs*, *21*(4), 272–279.

Shriberg, E. (1995). Acoustic Properties of Disfluent Repetitions. *Proceedings of International Conference on Phonetic Sciences (ICPhS)*, *4*, 384–387.

Šimko, J., & Beňuš, Š. (2016). Hyperarticulation in Lombard speech: Global coordination of the jaw, lips and the tongue. *Journal of the Acoustical Society of America*, *139*(1), 151–162.

Šimko, J., & Cummins, F. (2011). Sequencing and optimization within an embodied task dynamic model. *Cognitive Science*, *35*(3), 527–562. https://doi.org/10.1111/j.1551-6709.2010.01159.x

Smith, A. (1992). The Control of Orofacial Movements in Speech. *Critical Reviews in Oral Biology & Medicine*, *3*(3), 233–267. https://doi.org/10.1177/10454411920030030401

Smith, A. (2006). Speech motor development: Integrating muscles, movements, and linguistic units. *Journal of Communication Disorders*, *39*(5), 331–349. https://doi.org/10.1016/J.JCOMDIS.2006.06.017

Smith, A., & Goffman, L. (1998). Stability and Patterning of Speech Movement Sequences in Children and Adults. *Journal of Speech, Language, and Hearing Research*, *41*(1), 18. https://doi.org/10.1044/jslhr.4101.18

Smith, A., Goffman, L., Sasisekaran, J., & Weber-Fox, C. (2012). Language and motor abilities of preschool children who stutter: Evidence from behavioral and kinematic indices of nonword repetition performance. *Journal of Fluency Disorders*, *37*(4), 344–358. https://doi.org/10.1016/J.JFLUDIS.2012.06.001

Smith, A., Kelly, E., Curlee, R. F., & Siegel, G. M. (1997). Stuttering: A dynamic, multifactorial model. *Nature and Treatment of Stuttering: New Directions*, *2*, 204–217.

Smith, B. L., Sugarman, M. D., & Long, S. H. (1983). Experimental manipulation of speaking rate for studying temporal variability in children's speech. *The Journal of the Acoustical Society of America*, *74*(3), 744–749. https://doi.org/10.1121/1.389860

Sowman, P. F., Ryan, M., Johnson, B. W., Savage, G., Crain, S., Harrison, E., … Burianová, H. (2017). Grey matter volume differences in the left caudate nucleus of people who stutter. *Brain and Language*, *164*, 9–15. https://doi.org/10.1016/j.bandl.2016.08.009

Stager, S. V, & Ludlow, C. L. (1998). The effects of fluency-evoking conditions on voicing onset types in persons who do and do not stutter. *Journal of Communication Disorders*, *31*(1), 33–52. https://doi.org/10.1016/S0021-9924(97)00049-X

Starkweather, C. W., & Gottwald, S. R. (1990). The demands and capacities model II: Clinical applications. *Journal of Fluency Disorders*, *15*(3), 143–157. https://doi.org/10.1016/0094-730X(90)90015-K

Starkweather, C. W., Hirschman, P., & Tannenbaum, R. S. (1976). Latency of vocalization onset: Stutterers versus nonstutterers. *Journal of Speech, Language, and Hearing Research*, *19*(3), 481–492.

Starkweather, C. W., & Myers, M. (1979). Duration of subsegments within the intervocalic interval in stutterers and nonstutterers. *Journal of Fluency Disorders*, *4*(3), 205–214.

Steiner, I., Richmond, K., & Ouni, S. (2013). Speech animation using electromagnetic articulography as motion capture data. In *12th International Conference on Auditory-Visual Speech Processing* (pp. 55–60). Annecy, France.

Stetson, R. H. (1928). *Motor phonetics: A study of speech movements in articulation*. Oberlin College, Oberlin: Springer Netherlands. https://doi.org/https://doi.org/10.1007/978-94-015-3356-0

Stone, M. (1997). Laboratory techniques for investigating speech articulation. In W. J. Hardcastle, J. Laver, & F. E. Gibbon (Eds.), *The Handbook of Phonetic Sciences* (2nd ed., pp. 7–38). Wiley Blackwell Publishing Ltd.

Stone, M. (2004). *A guide to analysing tongue motion from ultrasound images*. *Tongue features from ultrasound images*.

Stone, M., & Lundberg, A. (1994). Tongue-palate interactions in consonants vs. vowels. In *International Converence on Spoken Language Processing (ICSLP)* (pp. 49–53). Yokohama, Japan.

Stromsta, C. (1986). *Elements of stuttering*. Oshterno, MI: Atsmorts Publishing.

Strycharczuk, P., & Scobbie, J. M. (2015). Velocity measures in ultrasound data. Gestural timing of post-vocalic /l/ in English. In *International Congress of Phonetic Sciences (ICPhS)* (Vol. 18). Glasgow.

Subramanian, A., Yairi, E., & Amir, O. (2003). Second formant transitions in fluent speech of persistent and recovered preschool children who stutter. *Journal of Communication Disorders*, *36*, 59–75.

Sussman, H. M., Byrd, C. T., & Guitar, B. E. (2011). The integrity of anticipatory coarticulation in fluent and non-fluent tokens of adults who stutter. *Clinical Linguistics & Phonetics*, *25*(3), 169–186. https://doi.org/10.3109/02699206.2010.517896

Sussman, H. M., Hoemeke, K. A., & Ahmed, F. S. (1993). A cross-linguistic investigation of locus equations as a phonetic descriptor for place of articulation, *94*(3), 1256–1268.

Sussman, H. M., Hoemeke, K. A., & McCaffrey, H. A. (1992). Locus equations as an index of coarticulation for place of articulation distinctions in children. *Journal of Speech, Language, and Hearing Research*, *35*(4), 769–781.

Sussman, H. M., & Shore, J. (1996). Locus equations as phonetic descriptors of consonantal place of articulation. *Perception and Psychophysics*, *58*(6), 936–946. https://doi.org/10.3758/BF03205495

Tasko, S. M., & Greilick, K. (2010). Acoustic and Articulatory Features of Diphthong Production: A Speech Clarity Study. *Journal of Speech, Language, and Hearing Research*, *53*(February), 84–99. https://doi.org/10.1044/1092-4388(2009/08-0124)

Tasko, S. M., & McClean, M. D. (2004). Variations in articulatory movement with changes in speech task. *Journal of Speech, Language, and Hearing Research*, *47*(1), 85–100.

Tasko, S. M., Mcclean, M. D., & Runyan, C. M. (2007). Speech motor correlates of treatment-related changes in stuttering severity and speech naturalness, *40*, 42–65. https://doi.org/10.1016/j.jcomdis.2006.04.002

Tasko, S. M., & Westbury, J. R. (2002). Defining and measuring speech movement events. *Journal of Speech, Language, and Hearing Research*, *45*(1), 127–142.

Teesson, K., Packman, A., & Onslow, M. (2003). The Lidcombe behavioral data language of stuttering. *Journal of Speech, Language, and Hearing Research*, *46*(August), 1009–1015. https://doi.org/10.1044/1092-4388(2003/078)

Terband, H., Maassen, B., Guenther, F. H., & Brumberg, J. (2009). Computational Neural Modeling of Speech Motor Control in Childhood Apraxia of Speech (CAS). *Journal of Speech Language and Hearing Research*, *52*(6), 1595. https://doi.org/10.1044/1092-4388(2009/07-0283)

Throneburg, R. N., Yairi, E., & Paden, E. P. (1994). Relation between phonological difficulty and the occurrence of disfluencies in the early-etage of stuttering. *Journal of Speech and Hearing Research*, *37*(3), 504–509.

Unser, M., & Stone, M. (1992). Automated detection of the tongue surface in sequences of ultrasound images. *Journal of Acoustical Society of America*, *91*(5), 3001–3007.

van Lieshout, P. H. H. M. (1995). *Motor planning and articulation in fluent speech of stutterers and nonstutterers*. Nijmegen.

van Lieshout, P. H. H. M. (2017). Coupling dynamics in speech gestures: amplitude and rate influences. *Experimental Brain Research*, *235*(8), 2495–2510. https://doi.org/10.1007/s00221-017-4983-7

van Lieshout, P. H. H. M., Hulstijn, W., & Peters, H. F. M. (1996). Speech Production in People Who Stutter: Testing the Motor Plan Assembly Hypothesis. *Journal of Speech, Language, and Hearing Research*, *39*(1), 76–92.

van Lieshout, P. H. H. M., & Moussa, W. (2000). The assessment of speech motor behavior using electromagnetic articulography. *The Phonetician*, *81*(1), 9–22.

van Lieshout, P. H. H. M., Namasivayam, A. K., & Maassen, B. (2010). Speech motor variability in people who stutter. In *Speech Motor Control: New Developments in Basic and Applied Research* (pp. 191–214). Oxford University Press. https://doi.org/10.1093/acprof

van Lieshout, P. H. H. M., Peters, H. F. M., Starkweather, C. W., & Hulstijn, W. (1993). Physiological differences between stutterers and nonstutterers in perceptually fluent speech. *Journal of Speech, Language, and Hearing Research*, *36*(1), 55. https://doi.org/10.1044/jshr.3601.55

van Riper, C. (1982). *The nature of stuttering*. Englewood Cliffs N.J.: Prentice Hall.

Vasic, N., & Wijnen, F. (2005). Stuttering as a monitoring deficit. In R. J. Hartsuiker, R. Bastiaanse, A. Postma, & F. Wijnen (Eds.), *Phonological encoding and monitoring in normal and pathological speech* (pp. 226–247). Hove: Psychology Press.

Villegas, J., Wilson, I., Iguro, Y., & Erickson, D. (2015). Effect of a fixed ultrasound probe on jaw movement during speech. In *Ultrafest VI*. Edinburgh, UK.

Volenec, V. (2015). Coarticulation. In J. Davis (Ed.), *Phonetics* (pp. 47–86). New York: Nova Science Publishers.

Walsh, B., & Smith, A. (2002). Articulatory Movements in Adolescents. *Journal of Speech, Language, and Hearing Research*, *45*(6), 1119–1133. https://doi.org/10.1044/1092-4388(2002/090)

Ward, D. (1997). Intrinsic and extrinsic timing in stutterers' speech: Data and implications. *Language and Speech*, *40*(3), 289–310. https://doi.org/10.1177/002383099704000305

Ward, D. (2018). *Stuttering and cluttering* (Second Edi). London; New York: Taylor & Francis.

Watanabe, M., Hirose, K., Den, Y., & Minematsu, N. (2008). Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners. *Speech Communication*, *50*(2), 81–94. https://doi.org/10.1016/J.SPECOM.2007.06.002

Watkins, M., Baptista, B., & Watkins, M. (2006). Variability in the use of weak forms of prepositions. In *Studies in Blingualism* (Vol. 31, pp. 171–183).

Watson, B. C., & Alfonso, P. J. (1982). A comparison of LRT and VOT values between stutterers and nonstutterers *. *Journal of Fluency Disorders*, *7*(2), 219–241.

Wenig, P., & Conrad, B. (1987). Electromagnetic Articulography : Use of Alternating Magnetic Fields for Tracking Movements of Multiple Points Inside and Outside the Vocal Tract, *35*, 26–35.

Wieneke, H. J., Eijken, E., Janssen, P., & Brutten, G. J. (2001). Durational variability in the fluent speech of stutterers and nonstutterers. *Journal of Fluency Disorders*, *26*, 43–53.

Wiethan, F., Ceron, M., Marchetti, P., Giacchini, V., & Mota, H. B. (2015). The use of electroglottography, electromyography, spectrography and ultrasound in speech research-theoretical review. *Revista CEFAC*, *17*, 115–125.

Willis, D. (2006). Negation in Middle Welsh. *Studia Celtica*, *40*(Williams 1935), 63–88.

Wingate, M. E. (1964). A Standard Definition of Stuttering. *Journal of Speech and Hearing Disorders*, *29*(4), 484–489. https://doi.org/10.1044/jshd.2904.484

Wingate, M. E. (1969a). Sound and Pattern in "Artificial" Fluency. *Journal of Speech Language and Hearing Research*, *12*(4), 677. https://doi.org/10.1044/jshr.1204.677

Wingate, M. E. (1969b). Stuttering as a phonetic transition defect. *Journal of Speech and Hearing Disorders*, *34*(1), 107–108. https://doi.org/10.1097/gme.0b013e3181967b88

Wingate, M. E. (1976). *Stuttering: Theory and treatment*. Irvington.

Wingate, M. E. (1982). Early position and stuttering occurrence. *Journal of Fluency Disorders*, *7*(2), 243–258. https://doi.org/10.1016/0094-730X(82)90011-0

Wingate, M. E. (1984a). Fluency, disfluency, dysfluency and stuttering. *Journal of Fluency Disorders*, *17*, 163–168.

Wingate, M. E. (1984b). Stutter events and linguistic stress. *Journal of Fluency Disorders*, *9*, 295–300.

Wingate, M. E. (1988). The Fault Line. In *The Structure of Stuttering: A Psycholinguistic Analysis* (pp. 179–185). New York, Berlin: Springer Science & Business Media.

Winter, B. (2013). Linear models and linear mixed effects models in R with linguistic applications: Tutorial 1. *ArXiv Preprint ArXiv:1308.5499*.

Winter, B. (2015). A very basic tutorial for performing linear mixed effects analyses: Tutorial 2. *ArXiv Preprint ArXiv:1308.5499*.

Wishart, J., Timmins, C., McCann, J., Hardcastle, W. J., & Wood, S. E. (2008). The potential role of electropalatography (EPG) in improving speech intelligibility in children with Down Syndrome. *Journal of Intellectual Disability Research*, *52*(8), 677.

Wolk, L., Edwards, M. L., & Conture, E. G. (1993). Coexistence of Stuttering and Disordered Phonology in Young Children. *Journal of Speech, Language, and Hearing Research*, *36*(5), 906. https://doi.org/10.1044/jshr.3605.906

Wood, S. E., Timmins, C., Wishart, J., Hardcastle, W. J., & Cleland, J. (2019). Use of electropalatography in the treatment of speech disorders in children with Down syndrome: a randomized controlled trial. *International Journal of Language & Communication Disorders*, *54*(2), 234–248. https://doi.org/10.1111/1460-6984.12407

Wood, S. E., Wishart, J., Hardcastle, W. J., Cleland, J., & Timmins, C. (2009). The use of electropalatography (EPG) in the assessment and treatment of motor speech disorders in children with Down's syndrome: Evidence from two case studies. *Developmental Neurorehabilitation*, *12*(2), 66–75. https://doi.org/10.1080/17518420902738193

Wrench, A. A. (2015). Articulate Assistant Advanced User Guide. *Edinburgh: Articulate Instruments Ltd*.

Wrench, A. A., Cleland, J., & Scobbie, J. M. (2011). An ultrasound protocol for comparing tongue contours: upright vs. supine. In *Proceedings of 17th International Clinical Phonetics and Linguistics Association (ICPLA)* (Vol. 17, pp. 2161–2164). Hong Kong, China.

Wrench, A. A., Gibbon, F. E., McNeill, A. M., & Wood, S. E. (2002). An EPG therapy protocol for remediation and assessment of articulation disorders. In *7th International Conference on Spoken Language Processing* (pp. 965–968). Denver, Colorado.

Wrench, A. A., & Scobbie, J. M. (2006). Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. In *International Seminar on Speech Production (ISSP)* (Vol. 7, pp. 451–458).

Wrench, A. A., & Scobbie, J. M. (2008). High-speed Cineloop Ultrasound vs. Video Ultrasound Tongue Imaging: Comparison of Front and Back Lingual Gesture Location and Relative Timing. In *Proceedings of the Eighth International Seminar on Speech Production (ISSP)* (Vol. 8, pp. 57–60).

Wrench, A. A., & Scobbie, J. M. (2011). Very high frame rate ultrasound tongue imaging. In *International seminar on speech production (ISSP)* (Vol. 9, pp. 155–162).

Wu, J. C., Maguire, C. A. G., Riley, G., Lee, A., Keator, D., Tang, C., … Najafi, A. (1997). Increased dopamine activity associated with stuttering Increased dopamine stuttering. *Clinical Neuroscience and Neuropsychology*, *8*, 767–770. https://doi.org/10.1097/00001756-199702100-00037

Xu, Y. (2017). *Syllable as a synchronization mechanism*. *ISCA Workshop on Experimental Linguistics*. London.

Yairi, E., & Ambrose, N. G. (2005). Early childhood stuttering. Austin, TX: Pro-Ed.

Yairi, E., & Ambrose, N. G. (2013). Epidemiology of stuttering: 21st century advances. *Journal of Fluency Disorders*, *38*(2), 66–87. https://doi.org/10.1016/j.jfludis.2012.11.002

Yaruss, J. S., & Conture, E. G. (1993). F2 transitions during sound/syllable repetitions of children who stutter and predictions of stuttering chronicity. *Journal of Speech, Language, and Hearing Research*, *36*(5), 883. https://doi.org/10.1044/jshr.3605.883

Yaruss, J. S., Newman, R. M., & Flora, T. (1999). Language and disfluency in nonstuttering children's conversational speech. *Journal of Fluency Disorders*, *24*(3), 185–207. https://doi.org/10.1016/S0094-730X(99)00009-1

Yaruss, J. S., & Quesal, R. W. (2006). Overall assessment of the speaker's experience of stuttering (OASES): Documenting multiple outcomes in stuttering treatment. *Journal of Fluency Disorders*, *31*(2), 90–115. https://doi.org/10.1016/j.jfludis.2006.02.002

Yunusova, Y., Green, J. R., & Mefferd, A. (2009). Accuracy assessment for AG500, electromagnetic articulograph. *Journal of Speech Language and Hearing Research*, *52*(2), 547. https://doi.org/10.1044/1092-4388(2008/07-0218)

Zebrowski, P. M., Conture, E. G., & Cudahy, E. A. (1985). Acoustic analysis of young stutterers' fluency: Preliminary observations. *Journal of Fluency Disorders*, *10*(3), 173–192. https://doi.org/10.1016/0094-730X(85)90009-9

Zharkova, N. (2013). A normative-speaker validation study of two indices developed to quantify tongue dorsum activity from midsagittal tongue shapes. *Clinical Linguistics & Phonetics*, *27*(6–7), 484–496. https://doi.org/10.3109/02699206.2013.778903

Zharkova, N. (2016). Ultrasound and acoustic analysis of sibilant fricatives in preadolescents and adults. *The Journal of the Acoustical Society of America*, *139*(5), 2342–2351. https://doi.org/10.1121/1.4947046

Zharkova, N., Gibbon, F. E., & Hardcastle, W. J. (2015). Quantifying lingual coarticulation using ultrasound imaging data collected with and without head stabilisation. *Clinical Linguistics & Phonetics*, *29*(4), 249–265. https://doi.org/10.3109/02699206.2015.1007528

Zharkova, N., & Hewlett, N. (2009). Measuring lingual coarticulation from midsagittal tongue contours: Description and example calculations using English /t/ and /ɑ/. *Journal of Phonetics*, *37*(2), 248–256. https://doi.org/10.1016/j.wocn.2008.10.005

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2008). An ultrasound study of lingual coarticulation in children and adults. In *International Seminar on Speech Production (ISSP)* (Vol. 8, pp. 161–164).

Zharkova, N., Hewlett, N., & Hardcastle, W. J. (2011). Coarticulation as an Indicator of Speech Motor Control Development in Children : An Ultrasound Study Acquisition of Coarticulation by Children. *Motor Control*, *15*, 118–140.

Zharkova, N., Hewlett, N., Hardcastle, W. J., & Lickley, R. J. (2014). Spatial and temporal lingual coarticulation and motor control in preadilescents. *Journal of Speech, Language, and Hearing Research*, *57*(2), 374–389. https://doi.org/10.1044/2014

Zimmermann, G. N. (1980). Articulatory dynamics of fluent utterances of stutterers and nonstutterers. *Journal of Speech, Language, and Hearing Research*, *23*(1), 95–107.

Zimmermann, G. N., Smith, A., & Hanley, J. M. (1981). Stuttering. *Journal of Speech, Language, and Hearing Research*, *24*(1), 25. https://doi.org/10.1044/jshr.2401.25

7 Appendices

7.1 Appendix A: Heyde, Scobbie & Finlayson (2015)

Searching for Closure: Seeing a Dip

Cornelia J Heyde[1], James M Scobbie[1], Ian Finlayson[1,2]

[1] Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, UK
[2] School of Philosophy, Psychology and Language Sciences (PPLS), Edinburgh University, Edinburgh, UK

Queen Margaret University — Clinical Audiology, Speech and Language Research Centre

Articulate Instruments

Objectives

- To define a reference space for quantitative kinematic analysis of the tongue
- To avoid the difficulties of defining a non-subjective external referent
- To specify referents that are data-intrinsic
- To allow for within and across speaker comparisons

Background

- Ultrasound offers valuable information on the active articulator's location and deformation
- Most approaches to quantitative analysis rely on external referents like the bite plane, the palate , the probe, or a fixed cranial location
- External referents are still subject to individual anatomical differences

Participants and Materials

- High speed ultrasound tongue imaging
- 9 speakers
- CV syllables with a voiceless velar stop /k/
  CV = /ki/ /ko/ /kə/
- Schwa preceding each CV syllable to control for the lingual starting position
- Minimum of 12 repetitions of each prompt, resulting in 36 to 40 tokens per speaker

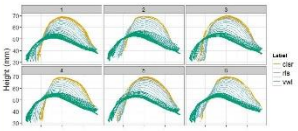


Figure 1: overlaid midsagittal tongue contour splines across six productions of the CV syllable /ka/ produced by speaker B

Recording & Postprocessing

- Data were recorded at 121fps using an Ultrasonix SonixRP running Articulate Assistant Advanced software
- Field of view 134.9°
- The probe was stabilised using a headset
- The probe was located medially to the chin bone and the hyoid bone to obtain maximal coverage of the tongue surface
- Acoustic data was segmented into
  CLSR closure,
  RLS release, and
  VWL subsequent vowel (Fig. 1)

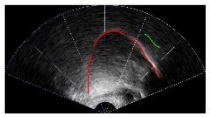


Figure 2: Single tongue contour spline at the acoustic midpoint of the closure (cf. yellow coloured splines in Fig. 1)

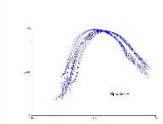


Figure 3: Overlaid tongue contour splines from 18 productions of three CV syllables

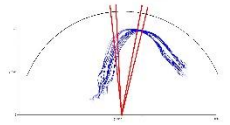


Figure 4: Two pairs of adjacent scanlines along which polar distances were measured and extracted for subsequent correlation (cf. Fig. 5)

Analysis

- Splines were fitted to the closure, release and subsequent vowel phases of the CV syllable (Fig. 2)
- Tongue contour splines at the acoustic midpoint of the closure were overlaid (Fig. 3)
- Polar distances from the probe to all tongue contour splines were extracted along 42 fanlines (3.2°)
- Each pair of adjacent distance sets was analysed using Pearson's r (Fig. 4 & 5) to measure the degree of correlation indicating a consistency of slope with which the splines cross through the fanlines

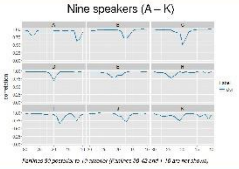


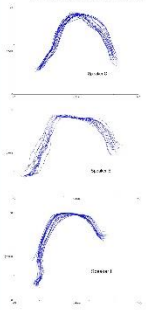
Figure 5: Pearson's r correlations of adjacent probe-to-spline radial distances, at the acoustic midpoint of the closure, ranging from posterior to anterior (left-to-right within each panel)

Observations

- Very high correlations of polar distances from adjacent fanlines
- A prominent dip in correlation (for all but one speaker) occurs roughly centrally in the fan (between fanline 17 and 23)
- We interpret this as the location of velar closure

Discussion

- The tongue surface is like an arch-shape, rather than being like the circumference of a circle
- The polar distances from the probe to posterior and anterior portions of the tongue surface therefore cross through adjacent fanlines at a consistently steep slope, resulting in high correlations
- Where the tongue surface is most circumferential relative to the fan, a stark drop in correlations is seen



Could dips be an indicator of coarticulation?

- Possibly, yes. However, additional factors appear to shape the dip:
- Speaker B (Fig. 3): clear distinction between fronted (/ki/) and non-fronted (/ka/ and /kə/) tongue contours causes a cross-over with low correlation
- Speaker C (left): equally prominent correlation dip (Fig. 5), although tongue contours do not show a strong binary fronted vs. non-fronted characteristic

Why does the width of the dip vary between speakers?

- It is positively related to the length of the tongue surface (i.e. number of fanlines) where the consistency of the slope with which the tongue crosses the fanlines is low (speaker E, left)

When do we observe multiple dips?

- Low correlations that are very posterior (fanline 30) or anterior (fanline 10) to the tongue surface (speaker K) are likely to be artefacts due to image noise
- Second prominent dips may also be caused by the tip of the tongue being circumferential to the fanlines (speaker I, left)

Conclusion

- Correlation dips appear to be systematic occurrences that are internal to the data and as such relatively speaker independent
- Dips mark the location of the most circumferential part of tongue, namely, the part of the tongue in palate contact, which we know from other work is the region where the tongue moves maximally
- Dips may be useful as a key to where a measurement vector for kinematic analysis should be located
- Measuring along the vector that crosses the dip will give a consistent and objective measure of displacement, velocity and duration of articulatory movement strokes

References:

[1] Articulate Instruments Ltd (2012). Articulate Assistant Advanced User Guide: Version 2.14. Edinburgh, UK: Articulate Instruments Ltd.
[2] Frisch, S. A. (2010). Semi-automatic measurement of stop-consonant articulation using edgetrak. Ultrafest V.
[3] Iskarous, K. (2005). Patterns of tongue movement. Journal of Phonetics, 33(4), 363-381.

Contact:
Cornelia J Heyde
cheyde@qmu.ac.uk

7.2 Appendix B: Heyde, Scobbie, Lickley & Drake (2016)

CLINICAL LINGUISTICS & PHONETICS
2016, VOL. 30, NOS. 3–5, 292–312
http://dx.doi.org/10.3109/02699206.2015.1100684

Taylor & Francis
Taylor & Francis Group

How fluent is the fluent speech of people who stutter? A new approach to measuring kinematics with ultrasound

Cornelia J. Heyde [a], James M. Scobbie [a], Robin Lickley [a] and Eleanor K. E. Drake [b]

[a]Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, Scotland; [b]School of Philosophy, Psychology and Language Sciences (PPLS), University of Edinburgh, Edinburgh, Scotland

ABSTRACT

We present a new approach to the investigation of dynamic ultrasound tongue imaging (UTI) data, applied here to analyse the subtle aspects of the fluency of people who stutter (PWS). Fluent productions of CV syllables (C = /k/; V = /ɑ, i, ə/) from three PWS and three control speakers (PNS) were analysed for duration and peak velocity relative to articulatory movement towards (onset) and away from (offset) the consonantal closure. The objective was to apply a replicable methodology for kinematic investigation to speech of PWS in order to test Wingate's Fault-Line hypothesis. As was hypothesised, results show comparable onset behaviours for both groups. Regarding offsets, groups differ in peak velocity. Results suggest that PWS do not struggle initiating consonantal closure (onset). In transition from consonantal closure into the vowel, however, groups appear to employ different strategies expressed in increased variation (PNS) versus decreased mean peak velocity (PWS).

ARTICLE HISTORY
Received 15 February 2015
Revised 18 June 2015
Accepted 22 September 2015

KEYWORDS
Dynamic analysis; fluency; gestural timing; stuttering; ultrasound

Introduction

Persistent developmental stuttering is a motor-speech disorder (Namasivayam & van Lieshout, 2011) which emerges in childhood. It is typically characterized by a relapsing-remitting, often situation-specific pattern of symptoms – primarily involuntary disruptions in the smooth flow of speech. These symptoms are described in terms of their acoustic consequences, labelled as blocks, prolongations and repetitions. The majority of the motor disruption underlying these acoustic consequences occurs within the (internal) vocal tract. It is therefore difficult to observe and measure the speech-motor activity directly involved in stuttering. For the same reason, it is usually difficult to compare the speech-motor performance during fluent speech of people who stutter (PWS) and those who do not, which is an important task if we hope to understand the sources of the disruptions. Ultrasound tongue imaging (UTI) offers a means to observe the speech-motor activity of the primary active oral articulator. It can therefore contribute meaningful additional information to the study of stuttering, particularly in light of the suggestion that stuttering is best understood as involving disruption to the high temporal coordination of oral (articulatory) and laryngeal (phonatory) movements (Adams, 1999; Max &

CONTACT Cornelia J. Heyde cheyde@qmu.ac.uk Speech and Hearing Sciences, Queen Margaret University, Queen Margaret Drive, Musselburgh, EH21 6UU, UK.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/iclp.

Gracco, 2005; Van Riper, 1982 for a review). In this article, we report a methodology we have adopted for investigating the dynamics of articulatory motor-speech production, both in PWS and in PNS. We will provide findings comparing the speech-motor productions of three PWS to three PNS using this ultrasound-based analysis.

Under experimental conditions, PWS perform more poorly across a range of acoustic measures of speech performance than do PNS. By their own rating and that of others, PWS are more susceptible to speech error elicitation than are PNS (Brocklehurst & Corley, 2011). PWS as a group also have longer speech reaction times (Cross & Luper, 1979; Harbison, Porter, & Tobey, 1989; Horii, 1984). Group differences between PWS and PNS in voice onset times (VOT) may be observable only in specific phonetic or utterance contexts (De Nil & Brutten, 1991; Healey & Ramig, 1986; Watson & Alfonso, 1982). As a group, PWS have been found to have longer vowel and consonant durations than PNS (Di Simoni, 1974; Starkweather & Myers, 1979). PWS were found to have descriptively longer VOT than PNS (Bakker & Brutten, 1990).

It is apparent that the poorer speech performance of PWS on acoustic measures reflects an underlying motor deficit of some nature. Between-group differences have been found for both non-speech and speech oro-motor performance (cf. Table 1). However, the fluctuating severity of stuttering symptoms indicates that the nature of the underlying motor deficit is probably complex and subtle: PWS are capable of producing speech that is acoustically indistinguishable from the speech of PNS. Articulatory performance has most commonly been assessed with reference to lip (L) and jaw (J) movement, as these articulators are the most accessible to observation. Early investigations into the relationship between phonatory and articulatory coordination employed photoglottographic recordings in conjunction with acoustic recordings (Yoshioka & Löfqvist, 1981). Subsequently the use of electroglottographic (EGG) and electromyographic (EMG) data from the lower lip allowed the calculation of physiological response times (as opposed to acoustic response times). Further EMG studies have revealed a general pattern of greater displacement and greater variability in lip movements in PWS than in PNS. This pattern is also apparent in studies employing either a strain gauge or a light-tracking approach, also measuring lower lip, upper lip and jaw (LL, UL and J) displacement (cf. Table 1). When a strain gauge approach has been used to investigate the sequencing of speech motor movements (for UL, LL, J), it has been found that atypical sequencing may be a consequence of adaptations rather than a primary symptom (McClean, Kroll, & Loftus, 1990).

Ultrasound tongue imaging

Ultrasound, like EMA, captures kinematic information about the key active oral articulator, namely the tongue. Another aspect that sets UTI and EMA apart from studies that investigate only the external articulators such as lips and jaw is that the tongue is crucial for most consonants and all vowels. But even though the tongue plays a role in consonants and vowels alike, the sequencing and overlap in time and space of different parts of the tongue need to be considered. UTI and EMA are not identical, however, in their suitability for providing such data. When measuring the kinematics of the tongue, EMA typically offers a better temporal and 2D spatial resolution than UTI. There are two aspects, however, where UTI is advantageous over EMA, namely that it provides holistic midsagittal tongue surface data, and that its output is not limited to just three or four anterior

Table 1. Studies investigating the speech and non-speech motor performance of people who stutter.

Study	Population	Instrumental approach	Topic investigated	Key findings
Chang, Ohde, & Conture (2002)	Children who stutter (CWS) v. children who do not stutter (CNS)	Acoustic measurement of formant transitions and F2 for CV syllables	Place of articulation and formant transitions	Groups differ in formant transition rate (FTR) as a function of place of articulation. CWS exhibit less contrast of FTRs between the labial and alveolar consonant contexts than CNS.
Namasivayam & van Lieshout (2008)	Adults who stutter (AWS) v. adults who do not stutter (ANS)	Electromagnetic articulography (EMA) Transducer coils on midline of vermilion border of upper and lower lips (UL, LL), lower jaw (J), the tongue blade (c. 1cm behind the anatomical tongue tip), the tongue body (c. 3cm behind tongue blade coil) and the tongue dorsum (c. 2cm behind tongue body coil). Only report bilabial productions.	Intergestural timing and stability	Amplitude of UL movement was significantly larger in PWS than PNS across normal and fast speech rates.
McClean, Tasko, & Runyan (2004)	AWS/ANS	EMA (UL, LL, TB, J)	Velocity, duration and speed ratios of different articulators	Complex pattern of findings: Task complexity interacted selectively with articulatory features
Smith, Sadagopan, Walsh, & Weber-Fox (2010)	AWS/ANS	Optotrak 3020 motion tracking system, tracking infrared light emitting diodes (IREDs) attached to the upper and lower lip (vermillion border). Tested nonword productions.	Articulatory stability	Higher lip aperture variability in AWS, especially in early trials compared to later trials.
Kleinow & Smith (2000)	AWS	IREDS attached to lower lip. Tested real words productions in carrier phrases.	Articulatory stability	Greater variability in AWS, who were also vulnerable to the phonological complexity of words whereas ANS were not.
Caruso, Abbs & Gracco (1988)	AWS/ANS	Strain gauge on UL, LL,J.	Inter-articulator sequencing	Between-group differences in the sequencing of movement onsets and velocity peaks
Max, Caruso, & Gracco (2003)	AWS/ANS	Speech, non-speech and finger movements. Tested real nouns with bilabial onsets, following 'my'. Used UL, LL, jaw strain gauge.		Between-group difference on lip and jaw closing (but not opening). AWS showed both longer movement durations and higher peak velocities and greater amplitudes during closing movements
Max & Gracco (2005)	AWS/ANS	EMA and EGG UL, LL, J and larynx	Inter-articulator sequencing	Longer acoustic durations for voice onset time and devoicing intervals for AWS. Group differences in kinematics of oral and laryngeal gesture coordination as measured by onset and peak velocity and vocal fold vibration (i.e. AWS show longer duration between laryngeal and oral onsets of movement)
Zimmermann (1980)	AWS/ANS	Cineradiography LL and jaw	Inter-articulator sequencing	Longer transition times and longer steady-state postures for AWS. Movements of AWS show greater asynchrony than those of ANS.

data points. (Also, UTI is more accessible.) In terms of spatial resolution, UTI is equivalent to EMA in radial directions relative to the probe (sub-millimetre accuracy), but is worse in circumferential measures, both as distance from the probe increases and as the number of echopulse beams within a given field of view decreases (Wrench & Scobbie, 2011). Both techniques are poor at imaging the tongue tip, since EMA's coils interfere with articulation, while UTI loses its capacity to image the tip if it is masked by the jaw shadow or raised to create a sublingual air pocket.

Regarding the nature of the kinematic measures, they therefore draw on different underlying spatiotemporal data. While UTI provides images of almost the entire tongue surface moving in time and space in a two-dimensional plane, EMA tracks the path of a few predetermined fleshpoints, typically but not necessarily in just two dimensions and just in the mid-sagittal plane. Typically for EMA, three or four electromagnetic coils are glued on the anterior part of the tongue's upper surface as close to a mid-sagittal site based on the tongue's symmetrical morphology as possible, and nowadays coils are recorded as they move in 3D, with analysis based on a data reduction to 2D movement within a cranial mid-sagittal plane. Ultrasound instead samples movement of the tongue's surface through a single plane and is typically orientated to cranial mid-sagittal orientation. It therefore captures an apparent mid-sagittal image of the tongue from near the tip right down to the root through space and time. This provides information not only about the tongue upper surface shape and location, but about tongue internal muscles (e.g. genioglossus), which can contribute to a principal component analysis. It is still regarded sufficient in most research to consider only the wealth of surface data which both techniques provide, in apparent 2D motion, while remembering the different nature of these idealisations. Since the tongue's midline and the cranial midline need not correspond exactly at rest, and since they vary during speech thanks to slight lateral asymmetries in speech production, the 2D data provided differ at source, even before we approach the holistic versus fleshpoint differences. Finally, of course, other crucial lateral and constrictional aspects of spatiotemporal production ought to be considered for a full picture, which requires using other techniques, such as electropalatography or MRI.

UTI is particularly relevant for clinical research, where we cannot know a priori where exactly to measure kinematics. We cannot know where the right place would be to place each EMA coil. The place of consonantal constriction may, for example, be more variable for experimental speakers with a speech disorder than for control speakers, and movement patterns of a coil in a suitable place for typical speech might be unrevealing for disordered speech. It is often not highlighted, in fact, that even for quantifying typical speech, the placing of an EMA coil is crucial, since slightly different coil placement provides a different kinematic trace and different analytic values. Greater study of how variation in EMA coil placement affects kinematic measures is needed in order to ensure the validity of data and derived measures. The same is of course true of kinematic measures from ultrasound, as we will see.

UTI is more easily accessible and less invasive than EMA. This point is particularly relevant when recruiting and testing clinical populations of relatively low incidence (e.g. at approximately 1% for stuttering, Craig, Hancock, Tran, Craig, & Peters, 2002), as UTI can be undertaken by a wider range of research teams and disciplines. The relatively noninvasive nature of UTI (compared to EMA) is valuable when working with

populations who may be particularly sensitive to and atypical in their adaptations to alterations in sensorimotor feedback, since EMA requires that people speak with wires emerging from between the lips (though obviously some speakers may not tolerate the headset needed to stabilise the UTI probe). The great advantage of EMA, however, is that the data from each coil are perfectly suited for dynamic analysis, and there is large literature of established techniques (Hoole & Nguyen, 1997; Schönle et al., 1987). On the other hand, quantitative analysis of UTI is typically static, in terms of the shape of the tongue at a segmental target. It usually relates the ultrasound data to associated acoustic events relative to which singular ultrasound frames are extracted (whether the acoustic midpoint, stop burst, or maximum constriction). Static UTI analysis has been employed to explore articulation from a variety of angles. Video-based ultrasound, with data output rates of 30 frames per second (which can be deinterlaced to 60 fps if appropriate), can, however, also be used for timing analysis. For many purposes, video rate output is as useful as high-speed ultrasound (Wrench & Scobbie, 2008), and it has been used to investigate socio-phonetic processes of timing (Lawson, Stuart-Smith, & Scobbie, 2014) and also processes of motor control (Zharkova, Hewlett, Hardcastle, & Lickley, 2014), including more specifically, coarticulation and intergestural timing (Gick & Campbell, 2003). Both experimental and theoretical evidence indicate that in order to investigate stuttering, it is valuable to explore temporal as well as spatial aspects of speech execution. Despite the optimism of Wrench and Scobbie (2008), the low number of frames used in video ultrasound probably limits such kinematic analyses too much. Not only may the few frames that are available not be able to meaningfully capture the more subtle nature of articulatory movement, but more importantly, a slower scan rate at the probe combined with buffering of data to create the images results in both temporal smearing of the raw image, double tongues and other spatial artefacts in the output images (Wrench & Scobbie, 2006). These make video data more suitable for analysis of the slow moving endpoints of articulatory-acoustic goals (i.e. the targets) than for kinematic analysis, especially of fast-moving articulations (Wrench & Scobbie, 2011). This is particularly important when investigating a disorder which essentially involves disruption to the smooth gestural flow of spoken output, where it is the process of articulatory-acoustic goal attainment which is of primary interest.

Dynamic analysis of ultrasound, to be similar to EMA, should therefore be based on a larger number of frames in the raw high-speed ultrasound data, for example, captured and stored at 120 frames per second or higher (Wrench & Scobbie, 2011). The large number of frames allows in-depth temporal and spatial investigation in principle. Articulatory events can be observed throughout the entire recording, enabling the researcher to explore events that are less predictable and not acoustically salient. Both aspects about temporal and spatial resolution of UTI are beneficial for detailed analysis of speech movements, even in qualitative analysis (Scobbie, Punnoose, & Khattab, 2013). Similar to EMA, measures of duration and velocity can be obtained, to shed light on the trajectory of the tongue surface and its components. This has been useful in the investigation of degree of coarticulation (Zharkova et al., 2014) and inter-gestural timing movement (Strycharczuk & Scobbie, 2015).

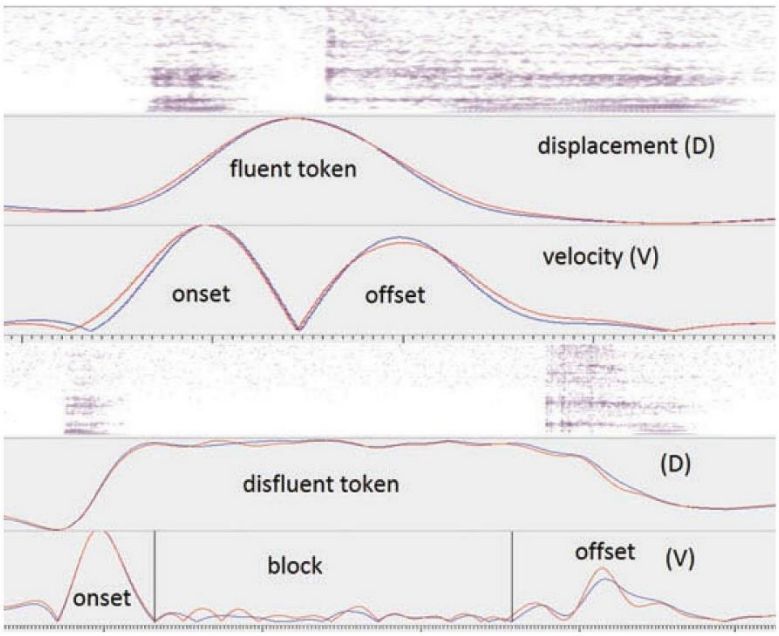


Figure 1. Example displacement and velocity traces for a fluent (top) and a disfluent (bottom) production of /ə kɑ/). Note that the disfluent production (1353ms) lasts approximately four times as long as the fluent production (327ms).

Fault-Line hypothesis

The theoretical framework for the present paper is based on Wingate's Fault-Line hypothesis (Wingate, 1988). The Fault-Line hypothesis responds to findings that PWS parse phrases based on syllables rather than utterances and that disfluencies typically commence on a consonant occurring on the first stressed syllable. Wingate claims that the main cause of disfluencies is the change in phonation (Wingate, 1976), which leads him to hypothesise that PWS do not struggle to initiate the consonant (i.e. the syllable onset) or the following vowel (nucleus), but to transition between them. The Fault-Line hypothesis (Wingate, 1988) therefore postulates that disfluencies result from PWS struggling to formulate rhymes (nucleus and coda) especially in stressed syllables. The underlying cause according to Wingate lies in the difficulty of retrieval and encoding of syllables rhymes which delay or inhibit the integration of the onset with its rhyme. With reference to the Fault-Line hypothesis, dynamic analysis of UTI can be of avail in the analysis of stuttering data in that it allows quantifying and comparing lingual coordination (i.e. duration and velocity). Looking at the different movement patterns in fluent and disfluent speech (Figure 1), it may be that the closing consonantal gesture for the velar constriction is indeed similar in both cases, with the difference between them being attributable entirely to the long block, where the tongue body perseveres in palatal constriction. Assuming an underlying motor impairment in PWS, differences should be observable also in their apparently fluent speech.

Aims

The purpose of the current study is to explore the potential of UTI as a tool to investigate motor coordination in the speech of PWS. In a pilot study, we explore syllable initial coordination in the fluent speech of three PWS and compare it to that of three control speakers. Sample data from CV syllables where C corresponds to the velar consonant /k/ is examined in detail. From the range of possibilities, one specific measurement vector is identified as appropriate for the quantitative analysis of one dimensional movement, and the apparent speed of the tongue surface up and down this vector is investigated and presented. Measures of displacement and velocity were collected at the point of maximum displacement of the tongue surface. Holistic movements are subdivided into movement 'strokes', which are derived from directly observable kinematics, which are then interpreted in terms of the underlying gesture. Each stroke is defined as the period between two successive minima in movement velocity along the measurement vector (Tasko & Westbury, 2002). Two movement strokes are of particular interest (Figure 1) – the movement towards the consonantal constriction of /k/ (i.e. the onset to closure) and the movement away from the constriction into a steady state of the vowel (i.e. the offset). In fluent speech, each stroke has a ballistic character, with a single peak velocity, reflecting aspects of the underlying gestural control parameters. Stroke durations and their peak velocities are compared within and between groups for three different vowel contexts following the /k/.

The hypotheses are that coordination patterns in the fluent speech of PWS and PNS will be found to behave similarly in duration and/or peak velocity for movement onset, whereas they will differ for offset movements. This would signify no gestural difficulty when initiating the absolute syllable-initial consonant and moving into its constriction, but a problem in either gestural planning or implementation when transitioning from the consonant into the following vowel.

Materials and methods

Participants

Three experimental speakers and three control speakers are reported (cf. Table 2 for demographic information). Speakers all fulfilled the two main criteria of fitting into the stabilising headset and providing good ultrasound image quality. The group of experimental participants self-reported as having a persistent developmental stutter (i.e. a stutter

Table 2. Demographic information for speakers.

	PWS 1	PNS 1	PWS 2	PNS 2	PWS 3	PNS 3
Gender	Female	Female	Male	Male	Male	Male
Age band	25–30	25–30	25–30	25–30	50–60	50–60
Handedness	Right	Right	Right	Right	Right	Right
Educational background	Post-graduate degree	Post-graduate degree	First degree	First degree	Information not provided	Post-graduate Degree
OASES	Moderate (58/253)		Mild-to-moderate (41/180)		Moderate-to-severe (62/253)	
SSI-IV	Moderate (26)		Mild (19)		Very severe (37)	

with an onset by age eight) (Büchel & Sommer, 2004; Prasse & Kikano, 2008). Stuttering severity was assessed using both a formal assessment (SSI-IV) and an assessment of the individuals' experience of their stutter (OASES). Results from both the SSI-IV (Riley, 2009) and OASES (Yaruss & Quesal, 2006) classified stuttering severity ranging between mild-to-moderate and moderate-to-severe. For all speakers, the last therapeutic intervention was a minimum of 5 years prior to recording. None of the speakers reported any lasting effect of the intervention. Speakers were recruited from the Edinburgh area. None of the participants reported any neurological, motor, auditory or visual impairment that could influence the outcome of the study. Speakers were compensated for their time with £15.

Stimuli

Data for the current study were part of a bigger corpus of recordings of adult speakers with a persistent developmental stutter. Target stimuli of the data presented are combinations of CV syllables with a voiceless velar stop (/k/) followed by a corner vowel (/i/ or /ɑ/) or schwa (/ə/). (It was decided that /u/ was too variable in placement in English to be used as a consistent context.) Each recording of a CV target item was preceded by a schwa (/ə/) to ensure a comparable lingual starting position. The preceding schwa was also useful in that it prevented bracing behaviours which would make it difficult to determine the time at which movement is initiated. Participants were instructed to stress the CV syllable following the schwa (i.e. to produce the pseudo-noun phrases 'a kaa', 'a kuh', 'a kee'). Fluent and disfluent recordings of the target stimuli /kɑ/, /kə/ and /ki/ will be presented separately.

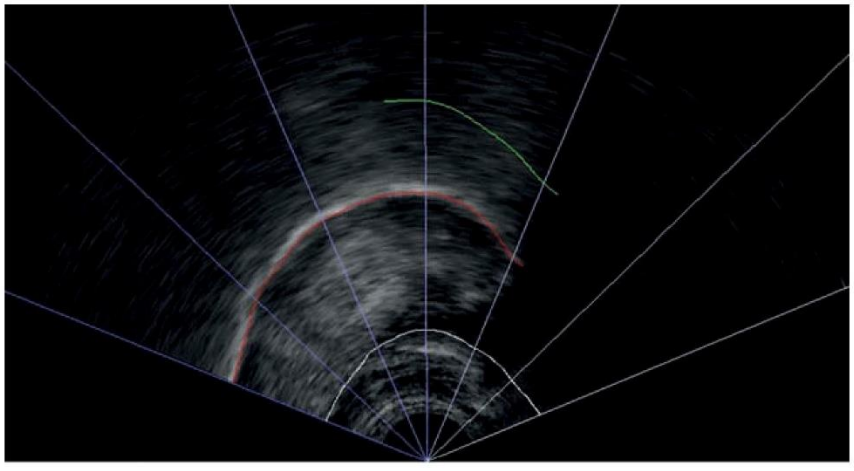


Figure 2. Ultrasound image showing the mid-sagittal tongue configuration (with tongue tip to the right and tongue root to the left) with an overlaid spline (red line) framed by traces of the hard palate (green line) and the floor of the mouth (white line).

Procedure

Participants were seated in front of a computer screen in a sound-treated recording room at Queen Margaret University. The ultrasound probe and a small microphone were attached to a stabilisation headset (Articulate Instruments Ltd., 2008) that participants wore during the recording session. The probe was oriented so as to display a mid-sagittal configuration with the tongue tip to the right and the root of the tongue on the left (Figure 2). The headset was used to control and reduce movement of the ultrasound probe as well as to ensure clarity of the ultrasound image. The ultrasound PC and a second control PC connected by Ethernet were located in a neighbouring control room, and data capture, synchronisation and data storage were controlled with Articulate Assistant Advanced software v2.14 (Articulate Instruments Ltd., 2012). The researcher in the control room initiated the beginning and the end of each token recording. As soon as the recording was initiated by the researcher, a fixation cross appeared on green background for 300 ms. Following this 300 ms delay, participants perceived a beep sound cueing them to read the prompt that appeared simultaneously on the screen. The ultrasound machine that was used to record the data was an Ultrasonix SonixRP, which has the advantage of being particularly precise in timing of ultrasound data capture and storage and in audio-visual synchronisation (Wrench & Scobbie, 2008). Data were recorded at ~121 frames per second (fps) with 63 echo pulse scan lines evenly spread over a 135-degree field of view (2.1° apart). The maximum depth was set to 80 mm and the echo return vectors had 412 samples resulting in a resolution of approximately 5 pixels per radial mm. The transducer frequency was 5 MHz, capable of resolving at a radial resolution of approximately 1 mm (Articulate Instruments Ltd., 2012). The data are stored in scanline format and reconstructed by AAA on the fly with radial interpolation to create a traditional fan-shaped image as the input to edge-fitting and subsequent analysis.

Fluency judgement

Recordings from PWS were categorised either as fluent or as disfluent. The categorisation was necessary to compare perceptually fluent recordings from PWS with those of PNS employing quantitative measures. Disfluent data were identified in two steps: The first author inspected the visual ultrasound and audible acoustic data of the entire corpus and extracted data that appeared acoustically and/or articulatorily aberrant from repetitions of the same target stimulus produced elsewhere by the same speaker. The preselected 'aberrant' recordings together with 'fluent' control versions of the same target stimulus were used for an objective evaluation. Twenty-five potentially disfluent recordings and 25 control recordings were randomised in an auditory judgement task. In a multiple forced-choice experiment, the recordings were presented in three randomised blocks (resulting in 150 stimuli overall) to five listeners. Listeners were trained linguists with no expertise in judging disfluent data. No hearing impairments were reported by listeners. After a brief introduction to the material, the recordings were presented to the listeners one by one, and they were instructed to judge whether the material appeared to be fluent or disfluent. Listeners were also required to indicate how certain they were about each judgement on a 4-point scale. Recordings were regarded as clearly disfluent when at least four out of the five listeners rated the recording as 'disfluent' with certainty (3 or 4 points on the 4-point scale).

Analysis

Dynamic data were analysed in four steps: (1) acoustic landmarking of word and segment locations, (2) splining of the tongue contour, (3) determination of location for measurement vectors and (4) kinematic annotations.

Acoustic landmarking

Acoustic data were exported from AAA into Praat (Boersma & Weenink, 2015) and semi-automatically annotated. An initial script distinguished silence from speech.[1] A following script opened each recording with the acoustic waveform and spectrogram. Automatically set boundaries were investigated and corrected when necessary. Additional boundaries were inserted distinguishing schwa, closure, release and vowel. Boundaries for vowels were based on periodic variation in the waveform preceding and following the voiceless consonant. The boundaries for the consonant were set at the onset of the stop consonant burst and at the onset of voicing for the following vowel. Acoustic landmark time-points were reimported into AAA and reintegrated with the audio/ultrasound data.

Splining of the tongue contour

Splines were inserted to track the edge of the tongue contour. Splines are mathematical functions that are useful for fitting a smooth curve to data. A spline is fitted to the shape of the tongue by defining a number of points along the length of the tongue: in AAA a default fan-shaped grid provides 42 equally-spaced control points over the whole fan-shaped image. For the first-pass analysis reported here, splines were fitted to the data on a reduced temporal sample rate of 40 splines per second (i.e. on every third frame of actual data), mainly for logistical reasons of time.

Semi-automated edge tracking was performed using AAA's built-in tracking functions on the frames displaying lingual movement from the schwa sound into the consonant /k/ and transitioning from /k/ onto the following vowel until the tongue contour reached a stable position following the release phase. An upper and a lower limit for automatic edge-fitting are specified for each frame in the same recording. To attach splines to each of the frames in a period of speech via the AAA tracking function, first an approximate tongue contour for the first frame is drawn manually. Its location is then refined semi-automatically with AAA's inbuilt 'snap-to-fit' function, a local search which scans along each of the 42 fan lines for the best dark-to-light edge. This function takes an input candidate spline and moves it to the clearest edge in its near neighbourhood, after which a built-in within-frame smoothing option can be applied to reduce radius-to-radius variation which can occur to the large number (up to 42) of closely-spaced knots (Articulate Instruments Ltd., 2012). This edge-tracking spline-fitting function therefore identifies the location of the tongue surface and gives it a relative confidence rating indicating the validity of the data at each knot. It is possible to manually correct the spline by moving incorrect knots, but particularly useful is the facility to set confidence to 0% at the anterior and posterior edges of the tongue surface in the image, which makes these irrelevant parts of the spline invisible to the user and to subsequent analysis.

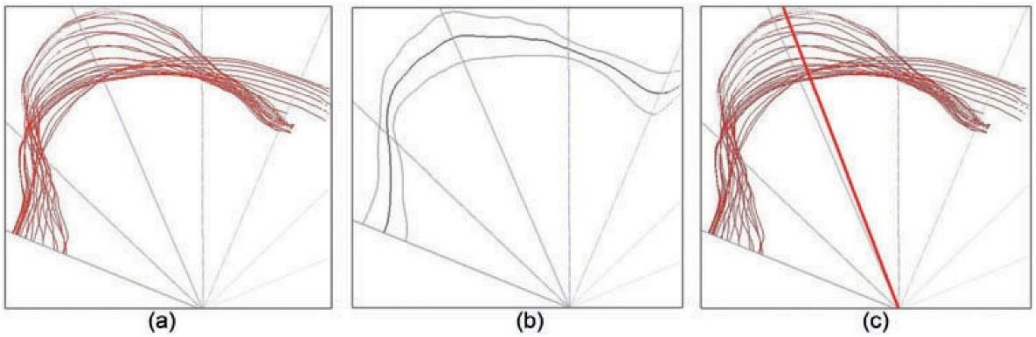


Figure 3. Superimposed 2D splines over time (a); confidence and mean standard deviations (b); measurement vector at selected fan line in velar area at maximum lingual displacement (c).

Measurement vector

The fitted first spline described above is the starting point for an automated edge tracking of the tongue surface contour throughout the subsequent frames of the recording. The tracking function bases a new spline in a frame on the shape already finalised for the preceding frame, then does a snap-to-fit local search for the new edge. The AAA tracking function re-iterates the snap-to-fit function automatically through a series of frames. Local search edge tracking works well for tracking dynamic changes in the mid-sagittal curve given a suitably high frame rate and clear images. These require a high underlying probe scan rate captured digitally because each frame is both a clear snapshot of a short time interval and only slightly different to the one preceding it.

Tracking the splines throughout the frames is based on a combination of edge detection and brightness detection. Tracking was interrupted and a new manual starting point for the splines was defined manually if artefacts in the ultrasound image led the automated tracking to go astray. To determine a suitable measurement vector for kinematic analysis, the splines that display onset, closure and offset of the tongue movement can be superimposed, to create a 2D image that informs about the extent of tongue movement at different points in the vocal tract, indirectly reflecting differential movement extent of different areas on the tongue surface (Iskarous, 2005). Splines are superimposed separately for each speaker and context because each vector needs to be not just speaker-specific but specific to the comparisons being made. The resulting image (Figure 3a) is used to estimate the strength of the signal as well as to establish areas on the tongue surface where displacement is largest. For the example given, investigating a velar consonant, the measurement vector should a priori be placed in the velar area: indeed this was where lingual displacement was largest. For all measurements presented here, the candidate vectors were all fan radii, given the instrumentally high resolution of ultrasound in radial directions and the nature of constrictions for /k/ in the vocal tract relative to the probe, though note that other orthogonal vectors could be used if thought necessary. The specific vector for analysis was chosen on objective criteria as follows. A mean spline was created based on the mean values for each spline-knot at each of the 42 fan radii, from the dynamic articulatory event of interest, with standard deviation and confidence indicated. All subsequent measurements were taken along the scan line with the greatest standard deviation in the velar area (Figure 3b), which was taken as indicative of the area of greatest movement. A relatively high

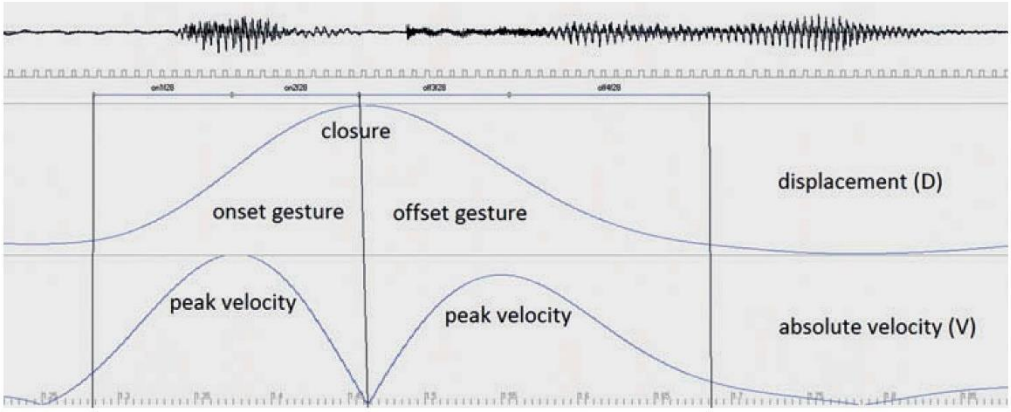


Figure 4. Displacement and velocity traces indicating onset (179 ms) and offset (224 ms) movement 'strokes' and the relative peak velocity for the CV syllable /ə kɑ/ lasting approximately 404 ms.

confidence of the semi-automatic AAA spline fitting (at least 85% overall) was used as a threshold for the validity of data. In different conditions, the radial line for which the lingual movement was largest (i.e. largest value of standard deviation from the mean spline) was adopted as the measurement vectors for kinematic analysis of tongue surface speed (cf. red line in Figure 3c).

Annotations

Both displacement and velocity were calculated by AAA for the spline as it changed location along the measurement vector. Displacement measures (D) indicate the radial distance from the origin of where a spline crosses the vector (Figure 4). Displacement and absolute velocity (V) data were captured for the relevant area including movement onset, lingual closure and the offset movement away from palatal constriction until a stable position for the vowel was reached. Based on the absolute velocity, two gestural 'strokes' could typically be readily identified for the production of the CV target stimulus. An inbuilt 'find function' (Articulate Instruments Ltd., 2012) was used to semi-automatically create annotations based on the absolute velocity profile (lower tier) starting from the point of zero velocity at acoustic closure. This point of zero velocity is preceded and followed by increases in velocity signifying the movement towards and away from consonantal constriction. Two regions were therefore identified based on the velocity trace for each recorded CV syllable, reaching back and forward in time from the stable target.

(1) Onset: from movement initiation away from a relatively stable position, via maximum velocity, up to consonantal constriction where absolute velocity reaches zero and
(2) Offset: from consonantal constriction with zero absolute velocity, via maximum velocity, until the tongue reaches a stable position for the vowel.

Measures for the beginning of the movement onset phase and the end of the offset phase are more arbitrary, ambiguous and sensitive to subtle but irrelevant articulatory movement, so were fixed at a time-point where the velocity exceeded a threshold of 20% of the

peak velocity when moving towards/away from the closure for the velar stop. This is a typical procedure in EMA studies (Pouplier & Waltl, 2008; Tasko & Westbury, 2002) and avoids undesirable and theoretically misleading variation in duration measurement.

Results

For reference, we include the results of statistical analyses on this pilot data set, but we caution that these analyses are most likely underpowered, particularly where speaker group differences are concerned (power analyses indicate a minimum requirement of eight participants per group in order to achieve a β-level of 0.08 at an α-level of 0.05). Results are reported where (i) they concern group differences and/or (ii) they reveal significant differences. The stability/variability of articulatory coordination was investigated in the fluent speech of PWS and PNS, which are shown in the graphs below. Only a few recordings of disfluent productions by PWS were available. The perceptual analysis resulted in 8.8% of PWS recordings (8 out of 91) being identified as clearly disfluent. Their comparison to the fluent data is secondary to the comparison of fluent speech between groups. Three distinct steps were undertaken: (1) Onset and offset movement durations across multiple repetitions of the syllable initial consonant /k/ were compared across vowel context for both groups. (2) Maximum velocities for onset and offset movements were examined again contrasting vowel context across the two groups of PWS and PNS. In order to verify differences in the lingual movement of PWS and control speakers, descriptive statistics were calculated for the durations and peak velocities for both groups. Median values together with standard deviation are reported for 195 recordings (total of 203 less 8 disfluent recordings) from six speakers (3 PWS and 3 PNS) distributed over three vowel contexts (i.e. /kɑ/ (n = 65), /ki/ (n =

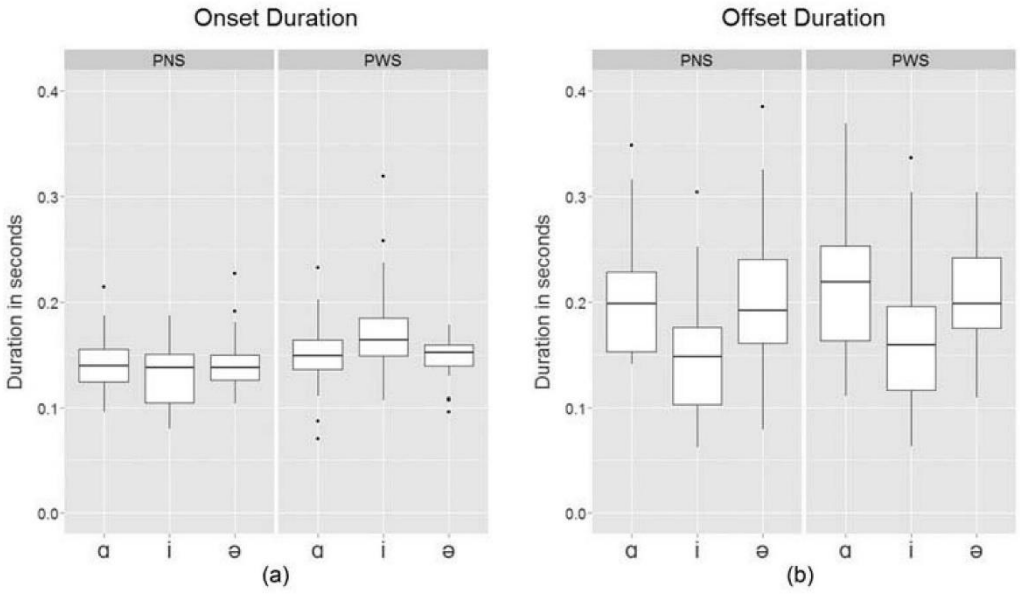


Figure 5. Duration measures for (a) onset and (b) offset by group (PNS vs. PWS) and vowel context (/ɑ, i, ə/).

Table 3. Mean duration measures (and standard deviations) in seconds for onset and offset strokes by vowel context and speaker group.

	All vowels		/ɑ/		/ə/		/i/	
	Onset	Offset	Onset	Offset	Onset	Offset	Onset	Offset
All	0.15	0.19	0.15	0.21	0.14	0.20	0.15	0.16
	(0.03)	(0.07)	(0.03)	(0.06)	(0.02)	(0.07)	(0.04)	(0.06)
PWS	0.16	0.20	0.15	0.21	0.15	0.20	0.17	0.17
	(0.04)	(0.07)	(0.03)	(0.06)	(0.02)	(0.05)	(0.04)	(0.07)
PNS	0.14	0.18	0.14	0.20	0.14	0.20	0.13	0.15
	(0.03)	(0.07)	(0.02)	(0.05)	(0.02)	(0.07)	(0.03)	(0.06)
PWS (disfluent)	0.22	0.24						
	(0.08)	(0.11)						

66) and /kə/ (n = 64). Each recording comprises an onset and an offset region relative to the gesture (see Figure 4).

Duration measures

Durations for onset and offset were calculated and are displayed in the two graphs representing the movement towards (Figure 5a) the velar /k/ closure and away from (Figure 5b) consonantal constriction for both groups. Data are presented for the three different vowel contexts (i.e. /ɑ/, /i/, /ə/) shown in the panels of the graphs.

Independent of vowel context and group, offsets are on average 0.04 s longer than onsets. While the duration for PWS increases from 0.16 s to 0.20 s, the duration for PNS increases from 0.14 s to 0.18 s. When the data are modelled, only slight duration differences between groups can be observed (onsets: β = 0.013 (SE = 0.009), t = 1.43; offsets: β = 0.016 (SE = 0.014), t = 1.225). Offset duration further shows increased variation when compared to onset duration with standard deviation increasing from approximately 0.03 up to 0.07 s. Durations for both onset and offset phases also tend to behave comparably across the three vowel contexts. Only the high vowel /i/ appears to affect offset movement duration, noticeably reducing it (/kɑ/ versus /ki/: β = −0.046 (SE=0.019), t = −2.453; /kɑ/ versus /kə/: β = −0.001(SE = 0.013), t = −0.110[2]; cf. Table 3 for descriptive data). While overall offset durations for both vowel contexts /ɑ/ and /ə/ are fairly constant (PWS: M 0.21/0.20 s; PNS: M 0.20 s/0.20 s), a decrease in offset duration (PWS: M 0.17 s; PNS: M 0.15 s) for /i/ vowel context is apparent.

Looking at data from only PWS and comparing the duration data for fluent (n of tokens = 83) with that of disfluent (n of tokens = 8) recordings, there is a noticeable difference in duration for onsets (fluent: M 0.16 s; disfluent: M 0.22 s) as well as offsets (fluent: M 0.20 s; disfluent: M 0.24 s). In the disfluent data, we find a similar pattern to that seen in the fluent data – shorter and less variable onsets compared to offsets (onset: M 0.22 ms, SD 0.08 s; offset: M 0.24 s, SD 0.11 s).

Peak velocity measures

Peak velocities for the tongue approaching the palate (i.e. onset velocity; Figure 6a) are comparable between groups (PWS: M 124 mm/s, SD = 52; PNS: M 126 mm/s, SD = 32 mm/s). In both speaker groups, mean peak velocity for offset movements (PWS: M 79 mm/s, SD =

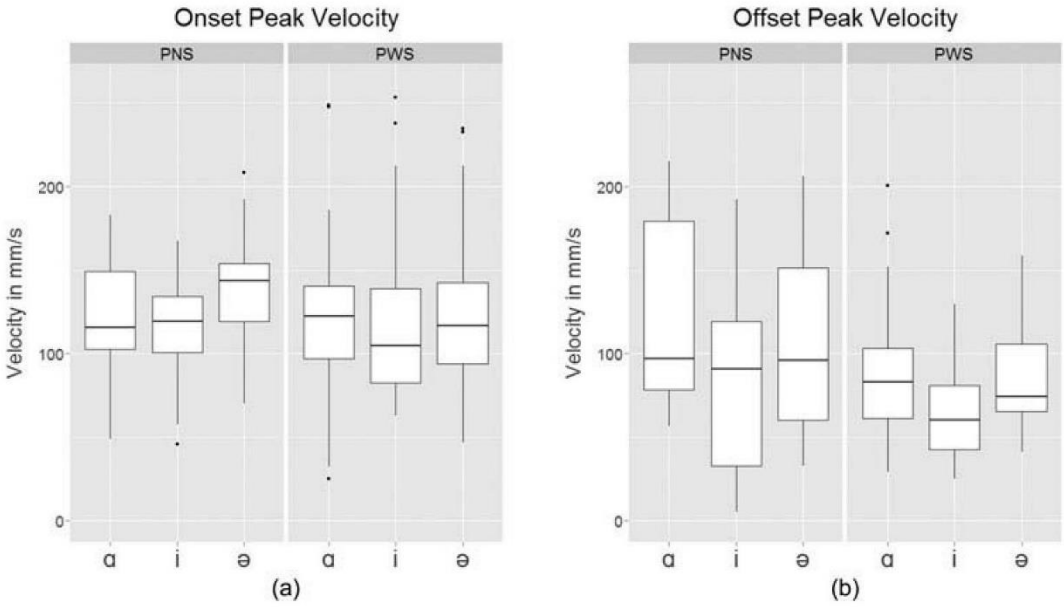


Figure 6. Peak velocity measures for (a) onset and (b) offset by group (PNS vs. PWS) and vowel context (/ɑ, i, ə/).

Table 4. Peak velocity measures (and standard deviations) in mm/s for onset and offset strokes by vowel context and speaker group.

	All vowels		/ɑ/		/ə/		/i/	
	Onset	Offset	Onset	Offset	Onset	Offset	Onset	Offset
All	125.08	93.83	122.50	109.38	133.25	97.59	119.70	74.87
	(41.85)	(48.36)	(41.78)	(51.46)	(40.96)	(43.49)	(42.18)	(43.82)
PWS	123.72	79.34	121.91	89.69	125.95	87.05	123.29	63.23
	(52.30)	(34.72)	(53.02)	(40.62)	(51.66)	(31.05)	(53.94)	(26.72)
PNS	126.09	104.40	122.90	122.00	138.57	105.50	116.70	84.24
	(32.21)	(54.01)	(33.02)	(54.13)	(30.66)	(49.84)	(29.55)	(52.33)

Mean peak velocity measures with standard deviations in brackets.

35 mm/s; PNS: *M* 104 mms/s, SD = 54 mm/s) is lower than for onset movements. This difference is stronger in PWS (Figure 6b). Statistically, group difference is not significant for either onsets (β = 13.39 (SE = 34.38), t = 0.390) or offsets (β = 14.06 (SE = 31.52), t = 0.446). However, it may be of theoretical significance that, as a group, PNS display greater variability in offset velocity than onset velocity, whereas for PWS the reverse is true.

Vowel context affects peak velocity for both groups (cf. Table 4). The vowel-dependent disparity in mean peak velocity is most evident for offset movements (/kɑ/ versus /ki/: β = −34.50 (SE = 17.95), t = −1.921; /kɑ/ versus /kə/: β = −12.04 (SE = 5.55), t = −2.169),[3] particularly when produced by PWS (with *M* 63 mm/s for /i/ compared to *M* 89 mm/s for /ɑ/ and *M* 87 mm/s for /ə/). Differences were also observable for onset velocity for /kɑ/ versus /kə/ (β = 11.103 (SE = 5.401), t = 2.056) but not for /kɑ/ versus /ki/ (β = 0.437 (SE = 9.074), t = 0.048).

In summary, offset (compared to onset) peak velocities are lower for PWS than PNS and display a larger vowel effect. For duration measures onset (when compared to offset) durations are shorter for both groups with overall slightly lower values for PNS than PWS

(fluent < disfluent recordings) (cf. Table 3). Patterns are consistent across vowel contexts with the exception of /i/.

Discussion

A novel approach to analysing dynamic ultrasound data was presented and applied to the fluent speech of speakers with a history of fluency problems (PWS). As predicted, PWS and controls (PNS) did not differ on duration or peak velocity measures when approaching an initial consonantal constriction (onset). Further, no statistically significant difference between groups was found for offsets. From this perspective, all speakers were equally fluent. The non-significance for between-group testing could result from (1) the nature of the stimuli or (2) the low number of speakers. Having included measures from only 'fluent' recordings, we by definition have eliminated the more apparent differences that could be expected otherwise (i.e. in the coordination of movements between fluent and disfluent speech). The low number of speakers makes larger values for variability more likely, reducing power to detect significant results. Results are therefore meaningful for descriptive analysis more so than for inferential statistics.

Vowel effects were observed for offsets affecting both duration and peak velocity measures. Effects observed for /i/ on offset duration measures may be due to an increased coarticulatory effect of /i/ onto /k/, decreasing the trajectory through the fronting of /k/. The vowel effects observed for offset peak velocity measures, on the other hand, may be due to an enhanced looping effect on /i/ in combination with /k/ (cf. Mooshammer, Hoole, & Kühnert, 1995). In both cases, forward movement would intersect the measurement vector (i.e. forward movement is essentially perpendicular to the measurement vector): it would not be measureable. However, it would most probably reduce the movement measured on the vector in these results. The more general vowel effects observed for velocity offsets further indicate a high degree of accuracy of the proposed method.

Referring to descriptive data, results support the notion that PWS do not struggle when moving towards the consonant, but do when transitioning from the consonant into the vowel. Both groups appear to use different strategies while still reaching the vowel target at the same time. Differences were observable for peak velocity measures in offsets, which are the transitions away from consonantal constriction towards a stable position in the following vowel. The descriptive statistics presented strongly suggest that PWS and PNS differed regarding both mean peak velocity and its degree of variation. Measures for PWS displayed observably lower means for peak velocity compared to those from control speakers. Control speakers on the other hand showed an overall larger variability of peak velocity. The lower overall peak velocity in offsets could suggest that PWS have fundamentally lower acceleration/deceleration compared to PNS.

Because experimental speakers were adults who have had their stutter since childhood, they are highly likely to have found strategies to maintain perceptually fluent speech, and the generally lower peak velocity could be just one. Despite the different histories of stuttering therapy, the shared strategy of lower peak velocity during release could reflect its efficacy to the user in overcoming struggles in maintaining fluency. Further data are required to verify the present results.

In accordance with Wingate's Fault-Line Hypothesis, these preliminary results show that PWS do not struggle when initiating the syllable initial consonant. In contrast, the differences between groups in peak velocity may indicate struggles PWS have when transitioning on to the following vowel. The differences could demonstrate difficulty forming and integrating syllable rhymes with their onsets. The fact that kinematic differences between groups show in apparently fluent speech could be an indicator for an underlying motor control impairment not limited to temporary disruptions in the speech flow.

Because differences between groups only affect velocity, but not duration measures, they could only be observed when looking at articulatory data. The approach we presented for kinematic ultrasound analysis relies crucially on an ability to observe dynamic movement of the tongue, a universal articulator, rather than just the lips or jaw. It allowed us to closely investigate the movement of the tongue towards and away from a consonantal constriction. The dynamic nature of the data made the trajectory of the tongue surface clearly observable and articulatory events could be analysed based on kinematics. In our approach, we have attempted to create a systematic method of ultrasound kinematics that (1) is replicable, (2) allows for inter-speaker comparison, (3) is modifiable for different places of articulation and (4) can be simply extended, e.g. to other non-radial vectors. Movement trajectories were broken down into 'strokes', which refer to transitions between two articulatory gestures. Breaking down trajectories into 'stroke' duration and 'stroke' peak velocity guaranteed a more in-depth examination eventually revealing kinematic differences between groups. The approach presented uses maximum displacement as a referent for measures of duration and peak velocity. Starting from kinematic movement patterns, measures respond and adapt to the individuals' articulatory setting and tongue surface trajectory. This aspect renders inter-speaker comparison legitimate. Moreover, because measures are based on maximum displacement relative to the place of articulation of a sound, our approach is not limited to a specific place of articulation, but can also be adapted for a variety of sounds with differing places of articulation.

Concluding remarks

UTI is an easily accessible instrument that is noninvasive and provides relatively high-quality images. Notwithstanding, a number of impediments need to be considered and overcome. Owing to the fact that ultrasound, particularly in the study of dynamic data, is fairly new in the research of speech movements, there is a lack of established analytical procedures tested in different laboratories from practical and statistical perspectives. Moreover, traditional UTI analysis relies heavily on 'eye-balling' in order to determine the area of interest on the tongue surface. 'Eye-balling' however requires extensive experience with data of that particular kind. A more replicable/reproducible approach to methodological improvement was set out, and with further enhancements this sort of approach could help motivate the development of ultrasound kinematics by the wider research community. To validate the proposed method, further testing should be conducted using larger data sets also including, for example, alveolar /t/ or fricative /s/ targets. Also, to test the reliability of the proposed method, data should be analysed across sessions of the same speaker. It is to be hoped that simultaneous vector-based UTI and fleshpoint-based EMA or replicated UTI/EMA data sets will test and verify kinematic ultrasound.

While we have seen that some kinematic UTI analysis can be done in the style of (or at least inspired by) well-established methods for the analysis of EMA data (i.e. adaptation of definitions of units of measure such as movement 'strokes'), others need to be defined anew. One of the 'to-be-defined' aspects bears on the general lack of a common referent that measures could be related to. This comes for free, if somewhat arbitrarily, with EMA, since a coil is glued to the articulators and not removed within a session. Ultrasound however provides a wealth of possible vectors for measurement within a session, and it is not clear how to solve the data-reduction problem. But both techniques face similar challenges when we consider how best to orientate and relate data across sessions and across participants. For ultrasound, we need to define an angle or referent line that is shared across speakers. For EMA, we need to define coil placements that are both optimal and shared.

While our approach was sensitive to different vowel contexts, it clearly needs to be run on larger amounts of data to be able to quantify its sensitivity to different consonantal targets. We predict that it will be necessary to make use of non-radial measurement vectors, for example. The difficulty lies in defining how to locate them in order to make data available for inter-speaker comparison. In sum, we presented an approach in which we defined the location of maximum radial displacement relative to the probe centre, so that the maximal dynamic variation could be used as a defining characteristic, across all speakers, for a common kinematic analysis.

Notes

1. The script settings were the following: minimum pitch 60 Hz, time steps: 0 s, silence threshold: −25 dB, minimum silent interval duration: 0.3 s, minimum sounding interval duration: 0.1 s.
2. Including the interaction between prompt type and speaker group did not improve model fit.
3. Again, including the interaction between prompt type and speaker group did not improve model fit.

Acknowledgments

We thank all participants who gave up their time to complete the experiment. Special thanks go to Alan Wrench, Patrycja Strycharczuk, Steve Cowen, Ian Finlayson and Jana Walzog for their constant support. We gratefully acknowledge the support of Queen Margaret University, particularly the Clinical Audiology, Speech and Language Research Centre.

Declaration of interest

The authors report no conflicts of interest.

ORCID

Cornelia J. Heyde http://orcid.org/0000-0003-4604-0385
James M. Scobbie http://orcid.org/0000-0003-4509-6782
Robin Lickley http://orcid.org/0000-0003-2583-5461
Eleanor K. E. Drake http://orcid.org/0000-0002-3475-7840

References

Adams, D. C. (1999). Methods for shape analysis of landmark data from articulated structures. *Evolutionary Ecology Research, 1*(8), 959–970.

Articulate Instruments Ltd. (2008). *Ultrasound stabilisation headset users manual: Revision 1.4.* Edinburgh, UK: Articulate Instruments Ltd.

Articulate Instruments Ltd. (2012). *Articulate assistant advanced user guide: Version 2.14.* Edinburgh, UK: Articulate Instruments Ltd.

Bakker, K., & Brutten, G. J. (1990). Speech-related reaction times of stutterers and nonstutterers diagnostic implications. *Journal of Speech and Hearing Disorders, 55*(2), 295–299.

Boersma, P., & Weenink, D. (2015). *Praat: Doing phonetics by computer: Version 5.4.05.* (retrieved 17 February 2015 ed.). http://www.praat.org/.

Brocklehurst, P. H., & Corley, M. (2011). Investigating the inner speech of people who stutter: Evidence for (and against) the covert repair hypothesis. *Journal of Communication Disorders, 44* (2), 246–260.

Büchel, C., & Sommer, M. (2004). What causes stuttering? *PLoS Biology, 2*(2), 159–163.

Caruso, A. J., Abbs, J. H., & Gracco, V. L. (1988). Kinematic analysis of multiple movement coordination during speech in stutterers. *Brain: A Journal of Neurology, 111*, 439–456.

Chang, S., Ohde, R. N., & Conture, E. G. (2002). Coarticulation and formant transition rate in young children who stutter. *Journal of Speech, Language and Hearing Research, 45*(4), 676–688.

Craig, A., Hancock, K., Tran, Y., Craig, M., & Peters, K. (2002). Epidemiology of stuttering in the community across the entire life span. *Journal of Speech, Language, and Hearing Research, 45*(6), 1097–1105.

Cross, D. E., & Luper, H. L. (1979). Voice reaction time of stuttering and nonstuttering children and adults. *Journal of Fluency Disorders, 4*(1), 59–77.

De Nil, L. F., & Brutten, G. (1991). Voice onset times of stuttering and nonstuttering children: The influence of externally and linguistically imposed time pressure. *Journal of Fluency Disorders, 16* (2), 143–158.

Di Simoni, F. G. (1974). Preliminary study of certain timing relationships in the speech of stutterers. *The Journal of the Acoustical Society of America, 56*(2), 695–696.

Gick, B., & Campbell, F. (2003). Intergestural timing in English /r/. *Proceedings of the 15th International Congress of Phonetic Sciences*, 1–4.

Harbison, D. C., Jr., Porter, R. J., Jr., & Tobey, E. A. (1989). Shadowed and simple reaction times in stutterers and nonstutterers. *The Journal of the Acoustical Society of America, 86*(4), 1277–1284.

Healey, E. C., & Ramig, P. R. (1986). Acoustic measures of stutterers' and nonstutterers' fluency in two speech contexts. *Journal of Speech, Language, and Hearing Research, 29*(3), 325–331.

Hoole, P., & Nguyen, N. (1997). Electromagnetic articulography in coarticulation research. *Forschungsberichte Des Instituts Für Phonetik Und Sprachliche Kommunikation Der Universität München, 35*, 177–184.

Horii, Y. (1984). Phonatory initiation, termination, and vocal frequency change reaction times of stutterers. *Journal of Fluency Disorders, 9*(2), 115–124.

Iskarous, K. (2005). Patterns of tongue movement. *Journal of Phonetics, 33*, 363–381.

Kleinow, J., & Smith, A. (2000). Influences of length and syntactic complexity on the speech motor stability of the fluent speech of adults who stutter. *Journal of Speech, Language and Hearing Research, 43*(2), 548–559.

Lawson, E., Stuart-Smith, J., & Scobbie, J. M. (2014). A mimicry study of adaptation towards socially-salient tongue shape variants. *University of Pennsylvania Working Papers in Linguistics, 20*(2), 99–110.

Max, L., Caruso, A. J., & Gracco, V. L. (2003). Kinematic analyses of speech, orofacial nonspeech, and finger movements in stuttering and nonstuttering adults. *Journal of Speech, Language, and Hearing Research, 46*(1), 215–232.

Max, L., & Gracco, V. L. (2005). Coordination of oral and laryngeal movements in the perceptually fluent speech of adults who stutter. *Journal of Speech, Language, and Hearing Research, 48*(3), 524–542.

McClean, M. D., Kroll, R. M., & Loftus, N. S. (1990). Kinematic analysis of lip closure in stutterers' fluent speech. *Journal of Speech, Language, and Hearing Research, 33*(4), 755–760.

McClean, M. D., Tasko, S. M., & Runyan, C. M. (2004). Orofacial movements associated with fluent speech in persons who stutter. *Journal of Speech, Language, and Hearing Research, 47*(2), 294–303.

Mooshammer, C., Hoole, P., & Kühnert, B. (1995). On loops. *Journal of Phonetics, 23*(1), 3–21.

Namasivayam, A. K., & van Lieshout, P. (2008). Investigating speech motor practice and learning in people who stutter. *Journal of Fluency Disorders, 33*(1), 32–51.

Namasivayam, A. K., & van Lieshout, P. (2011). Speech motor skill and stuttering. *Journal of Motor Behavior, 43*(6), 477–489.

Pouplier, M., & Waltl, S. (2008). Articulatory timing of coproduced gestures and its implications for models of speech production. *Proceedings of the 8th International Seminar on Speech Production*, 19–22.

Prasse, J. E., & Kikano, G. E. (2008). Stuttering: An overview. *American Family Physician, 77*(9), 1271–1276.

Riley, G. D. (2009). Stuttering Severity Instrument: SSI-4. Austin, TX: Pro-Ed.

Schönle, P. W., Gräbe, K., Wenig, P., Höhne, J., Schrader, J., & Conrad, B. (1987). Electromagnetic articulography: Use of alternating magnetic fields for tracking movements of multiple points inside and outside the vocal tract. *Brain and Language, 31*(1), 26–35.

Scobbie, J. A., Punnoose, R., & Khattab, G. (2013). Articulating five liquids: a single speaker ultrasound study of Malayam. In L. Spreafico and A. Vietti (Eds.), *Rhotics: New data and perspectives* (pp. 99–124). Bozen-Bolzano: BU Press.

Smith, A., Sadagopan, N., Walsh, B., & Weber-Fox, C. (2010). Increasing phonological complexity reveals heightened instability in inter-articulatory coordination in adults who stutter. *Journal of Fluency Disorders, 35*(1), 1–18.

Starkweather, C. W., & Myers, M. (1979). Duration of subsegments within the intervocalic interval in stutterers and nonstutterers. *Journal of Fluency Disorders, 4*(3), 205–214.

Strycharczuk, P., & Scobbie, J. M. (2015). Velocity measures in ultrasound data: Gestural timing of post-vocalic /l/ in English. *Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS)*, Glasgow, UK.

Tasko, S. M., & Westbury, J. R. (2002). Defining and measuring speech movement events. *Journal of Speech, Language, and Hearing Research, 45*(1), 127–142.

Van Riper, C. (1982). *The nature of stuttering.* Englewood Cliffs, NJ: Prentice Hall.

Watson, B. C., & Alfonso, P. J. (1982). A comparison of LRT and VOT values between stutterers and nonstutterers. *Journal of Fluency Disorders, 7*(2), 219–241.

Wingate, M. E. (1976). *Stuttering: Theory and treatment.* Boston, MA: Irvington.

Wingate, M. E. (1988). *The structure of stuttering: A psycholinguistic analysis.* New York, NY: Springer Verlag.

Wrench, A. A., & Scobbie, J. M. (2006). Spatio-temporal inaccuracies of video-based ultrasound images of the tongue. *Proceedings of the 7th International Seminar on Speech Production*, 451–458.

Wrench, A. A., & Scobbie, J. M. (2008). High-speed cineloop ultrasound vs. video ultrasound tongue imaging: Comparison of front and back lingual gesture location and relative timing. *Proceedings of the Eighth International Seminar on Speech Production (ISSP)*, 57–60.

Wrench, A. A., & Scobbie, J. M. (2011). Very high frame rate ultrasound tongue imaging. *Proceedings of the 9th International Seminar on Speech Production (ISSP)*, 155–162.

Yaruss, J. S., & Quesal, R. W. (2006). Overall assessment of the speaker's experience of stuttering (OASES): Documenting multiple outcomes in stuttering treatment. *Journal of Fluency Disorders, 31*(2), 90–115.

Yoshioka, H., & Löfqvist, A. (1981). Laryngeal involvement in stuttering. *Folia Phoniatrica Et Logopaedica, 33*(6), 348–357.

Zharkova, N., Hewlett, N., Hardcastle, W. J., & Lickley, R. J. (2014). Spatial and temporal lingual coarticulation and motor control in preadolescents. *Journal of Speech, Language, and Hearing Research*, *57*(2), 374–388.

Zimmermann, G. (1980). Articulatory dynamics of fluent utterances of stutterers and nonstutterers. *Journal of Speech, Language, and Hearing Research*, *23*(1), 95–107.

Wenn Stotterer nicht stottern. Quantifizierung dynamischer Ultraschalldaten

Cornelia J. Heyde [1], *James M. Scobbie* [1]

Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, UK

cheyde@qmu.ac.uk, jscobbie@qmu.ac.uk

Abstract

Stottern wird traditionell akustisch definiert, wobei man sich oftmals ausschließlich auf akustisch-perzeptive Unterbrechungen im ansonsten flüssigen Redefluss stützt. Die detaillierte artikulatorische Analyse der flüssigen Sprache von Stotterern soll Auskunft darüber geben, ob Elemente des Stotterns eventuell selbst dann nachweisbar sind, wenn sie akustisch nicht wahrnehmbar sind. Eine weitere Frage, die wir mit der dynamischen Analyse der Ultraschalldaten beantworten wollen, ist, wo genau sich Stottern manifestiert. Mit Hinblick auf die Fault-Line Hypothese von Wingate [1] untersuchen wir insbesondere die Bewegung der Zunge in die Verschlußstellung (Anglitt) und vergleichen diese mit der Bewegung im Übergang von der Verschlußstellung zum darauffolgenden Vokal (Abglitt). Die Ergebnisse unserer Untersuchung deuten an, dass sich Stotterer selbst in der scheinbar flüssigen Sprache von Kontrollsprechern unterscheiden. Artikulatorische Unterschiede zwischen den beiden Sprechergruppen wurden im Übergang von Konsonant zu Vokal beobachtet, was die Annahme Wingate's bestärkt, dass beim Stottern das Problem nicht auf einem bestimmten Wort oder Laut, sondern im Übergang von einem zum nächsten Segment, liegt.

Schlüsselbegriffe: Artikulation, , Motorkontrolle, Stottern, Ultraschall, Kinematik

1. Einleitung

Stottern ist eine Unterbrechung im Redefluss und wird häufig in drei Hauptsymptome unterteilt: Wiederholungen (/k-k-kafe/), Verlängerungen (/ʃ:::u:/) und Blockaden (/k---ɔp͡f/) [2]. Hier stellt sich die Frage, ob das Problem in der Realisierung des initialen Konsonanten selbst oder, wie von Wingate [1] vorgeschlagen, im Übergang von Konsonant zu darauf folgendem Vokal besteht. Wingate vertritt die Meinung, dass Unterbrechungen im Redefluss auftreten, wenn der Sprecher den auf den Konsonanten folgenden Vokal nicht problemlos integrieren kann. Die drei Hauptsymptome des Stotterns stellen jeweils akustische Ereignisse dar, die in Form von Unterbrechungen im ansonsten nicht-auffälligen Sprachfluss wahrgenommen werden. Hier stellt sich die Frage, ob Stottern wirklich, wie weithin angenommen [3], die Artikulation lediglich kurzzeitig lokal beeinflusst, oder ob es sich um eine motorische Störung handelt, die sich generell auf die Artikulation der Betroffenen auswirkt und somit die akustisch-perzeptiven Stottereregnisse eventuell lediglich die „Spitze des Eisberges" darstellen.

Diese beiden Fragen nach der Reichweite (i.e., lokal vs. global) und der Lokalisierung von Stottereregnissen (auf einem Segment oder im Übergang von einem zum nächsten Segment) wurden anhand von akustisch unauffälligen Ultraschalldaten untersucht. So wurde die flüssige Sprache von Stotterern mit der von Kontrollsprechern auf kinematische Unterschiede hin untersucht. Zwei Bewegungsmuster wurden hierfür quantifiziert [4] – zum einen die Bewegung in die Verschlußstellung (Anglitt) und zum anderen die Bewegung im Übergang von der Verschlußstellung zum darauffolgenden Vokal (Abglitt). Für beide Bewegungsverläufe wurden Parameter der Dauer und der erreichten Maximalgeschwindigkeit erhoben. Anschließend wurden die Ergebnisse der stotternden Erwachsenen mit denen der Kontrollsprecher (KS) verglichen. Wingate's „Fault-Line" Hypothese ließ Unterschiede zwischen beiden Sprechergruppen im Übergang von Konsonant zu Vokal (Abglitt) vermuten, wohingegen im Anglitt keine Unterschiede erwartet wurden. Die Quantifizierung der kinematischen Ultraschalldaten erlaubt es auch feinere Unterschiede zwischen den Sprechergruppen statistisch zu analysieren und auszuwerten.

2. Methode

2.1. Teilnehmer

Daten von neun stotternden Erwachsenen und neun Kontrollsprechern wurden erhoben. Die Stotterer waren mindestens 18 Jahre alt und berichteten alle von persistierendem idiopathischem Stottern, welches vor dem achten Lebensjahr begonnen hat. Alle Teilnehmer wurden zudem offiziell diagnostiziert. Zwei standardisierte Tests („*Stuttering Severity Instrument*" und „*Overall Assessment of Speaker's Experience of Stuttering*") gaben Auskunft über die Schwere (von mild bis stark ausgeprägt) des Stotterns zum Zeitpunkt der Datenerhebung.

Die Kontrollsprecher wurden bestmöglich mit den stotternden Erwachsenen bezüglich des Alters (20 bis 60 Jahre; Durchschnittsalter: SE: 34,4; KS: 33,6, Standardabweichung: SE: 14,2; KS: 12,2), des Geschlechts (SE und KS: 6 männliche und 3 weibliche Sprecher), der Bildung (höchster Bildungsabschluss) und der Händigkeit (Rechtshänder mit der Ausnahme eines Teilnehmers), sowie Muttersprache (Britisches Englisch) abgestimmt. Keiner der Teilnehmer berichtete über neurologische, Hör-, Seh- oder andere Störungen, die die Ergebnisse der Studie hätten beeinträchtigen können.

2.2. Datenerhebung

Audiosignale und kinematische Ultraschalldaten wurden parallel erhoben. Die Daten bestanden aus einsilbigen CV Silben mit velarem konsonantischem Silbenkopf /k/ und variierendem vokalischen Silbenkern. Der Silbenkern bestand aus den beiden Kardinalvokalen /a/ und /i/, sowie dem Zentralvokal Schwa. Jeder Silbenproduktion ging ein Schwa-Laut voraus, der es uns ermöglichte, den Anglitt (Artikulation hin zum konsonantischen Verschlusslaut /k/) für die kinematische Analyse

zugänglich zu machen. Die Daten wurden in neun Listen randomisiert und die gleichen Listen wurden von beiden Sprechergruppen produziert. Pro Sprecher wurden 12 Wiederholungen jeder der drei Silbenkombination (/kɑ/, /ki/, /kə/) aufgenommen. In einer Perzeptionsstudie wurden die Daten, die eindeutig für unflüssig (gestottert) befunden wurden, von der weiteren Analyse ausgeschlossen. Insgesamt wurden 637 Silbenproduktionen von allen 18 Sprechern analysiert.

2.3. Datenanalyse

Die Ultraschalldaten wurden mithilfe der AAA Software [5] mit 120 Bildern pro Sekunde aufgenommen und bearbeitet. Splines (mathematische Kurven) wurden der Zungenkontur mittels Edge-Tracking nachempfunden und in die Ultraschallbilder eingezeichnet (*Abbildung 1a*). Die Splines einer CV Silbe wurden überlagert um die Koordinierung der Zungenoberfläche im zweidimensionalen Raum darzustellen (*Abbildung 1b*) und ein radialer Vektor wurde in der Region fixiert, in der die größte radiale Verschiebung (ausgehend von der Ultraschallsonde unterhalb des Kinns) der Zungenoberfläche beobachtet wurde (*Abbildung 1c*). Von der Verschiebung der Zungenoberfläche entlang des Vektors wurden zwei Verläufe abgeleitet (*Abbildung 2*): Die positive Verschiebung (mit größer werdendem Abstand zur Ultraschallsonde) wird in der Folge als „Anglitt" bezeichnet. Sie stellt die Bewegung der Zungenoberfläche hin zum konsonantischen Verschluss dar. Die negative Verschiebung (mit kleiner werdendem Abstand zur Ultraschallsonde) wird in der Folge als „Abglitt" bezeichnet. Sie stellt die Bewegung der Zungenoberfläche vom konsonantischen Verschluss hin zur vokalen Öffnung dar. Der Anfangspunkt des Anglitts, wie auch der Endpunkt des Abglitts wurden (in Anlehnung an die Methode vieler EMA Studien) mit einem Grenzwert von 20% der Maximalgeschwindigkeit bestimmt.

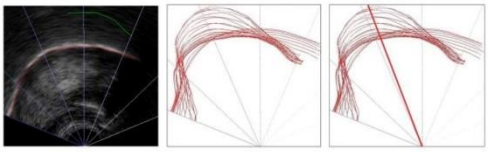


Abbildung 1: a) Ultraschallbild (Zungenspitze rechts; Zungenwurzel links) mit Spline (rote Kurve); b) überlagerte Splines relativ zu Anglitt und Abglitt; c) Fixierung des radialen Meßvektors (rote Linie) in der velaren Region

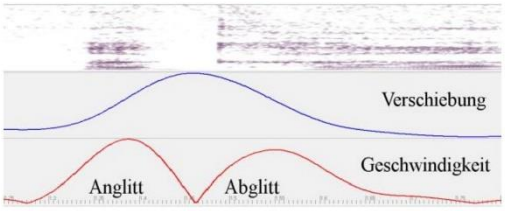


Abbildung 2: Artikulatorische Messungen von Dauer und Geschwindigkeit der Zungenoberfläche in der positiven (Anglitt) und negativen Verschiebung (Abglitt) entlang des Meßvektors für den Stimulus /ə ka/

Durchschnittswerte für die Dauer von Anglitt und Abglitt, sowie die Maximalgeschwindigkeit, die in Anglitt und Abglitt erreicht wurden, wurden für beide Sprechergruppen erhoben und verglichen.

3. Ergebnisse

Linear Paneldatenmodelle (mixed-effects models) wurden für die Analyse der Daten verwendet. Die Dauer (*Abbildung 3*) wie auch die Maximalgeschwindigkeit (*Abbildung 4*) wurden jeweils für An- und Abglitt auf signifikante Gruppeneffekte hin untersucht.

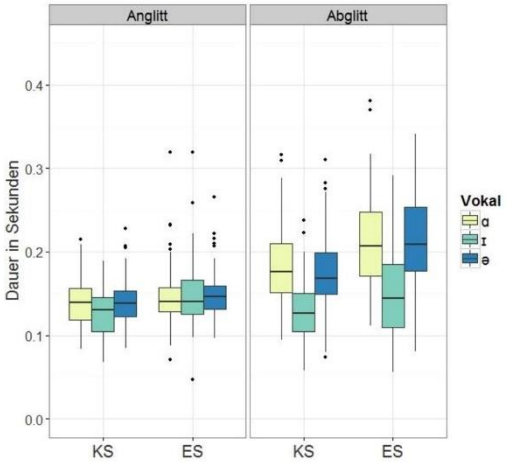


Abbildung 3: Durchschnittsdauer von Anglitt (Bewegung in die Verschlußstellung) und Abglitt (Bewegung im Übergang von der Verschlußstellung zum darauffolgenden Vokal) in Abhängigkeit von Sprechergruppe (Erwachsene Stotterer/ES im Vergleich zu den Kontrollsprechern/KS) und Vokal (/ɑ/, /i/, /ə/)

Der beste Bestimmungskoeffizient (Model-Fit) für die Analyse der Dauer des Anglitts (*Abbildung 3* – linke Spalte) beinhaltete Sprechergruppe (ES, KS) als Fixed-Effekt und Vokal des Silbenkerns (/kɑ/, /kə/, /ki/) in Abhängigkeit von Individuum als Random-Effekt. Dem Model zufolge gab es keinen signifikanten Unterschied zwischen den beiden Sprechergruppen in der Dauer des Anglitts (β = -0.008, SE = 0.007, t = 1.130). Für die Dauer des Abglitts (*Abbildung 3* – rechte Spalte) hat das Hinzufügen von Vokal als Fixed-Effekt den Model-Fit signifikant verbessert ($\chi^2(2) = 0.980$, p = 0.613). Der beste Model-Fit enthielt somit Sprechergruppe (SE, KS), wie auch Vokal (/ɑ/, /i/, /ə/) als Fixed-Effekt wobei die Steigung für den Vokal je Sprecher variieren konnte (Random-Effekt). Der Abglitt von stotternden Erwachsenen dauerte demnach signifikant länger als der der Kontrollsprecher (β = 0.030, SE = 0.009, t = 3.143). Der Abglitt hin zum Kardinalvokal /i/ war im Durchschnitt signifikant kürzer als der zum Vokal /a/ (β = -0.056, SE = 0.009, t = -6.423), wohingegen sich der Abglitt zum Schwa nicht von dem hin zum /a/ unterschied (β = -0.005, SE = 0.007, t = -0.678). Das Hinzufügen der Interaktion von Sprechergruppe und Vokal hat den Model-Fit nicht verbessert.

Die Maximalgeschwindigkeit von Anglitt und Abglitt verhielt sich ähnlich wie die Dauer: Während im Anglitt (*Abbildung 4* – linke Spalte) kein Unterschied zwischen den Sprechergruppen beobachtet wurde (β = -4.482, SE = 11.684, t = -0.384 mit Sprechergruppe als Fixed-Effekt während die Steigung für den Vokal je Sprecher variieren konnte), wurde im Abglitt (*Abbildung 4* - rechte Spalte) ein Effekt von Sprechergruppe (β = -24.576, SE = 9.923, t = -2.477), sowie von Vokal des Silbenkerns beobachtet. Der beste Model-Fit beinhaltete Sprechergruppe und Vokal als Fixed-Effekt mit Vokal in Abhängigkeit von Individuum als Random-Effekt.

Mit Bezug auf die Maximalgeschwindigkeit, die während des Abglitts erreicht wurde, wurde eine signifikant geringere Maximalgeschwindigkeit hin zum Vokal /i/ im Vergleich zum Vokal /a/ erreicht (β = -38.167, SE = 9.704, t = -3.933). Die erreichte Maximalgeschwindigkeit im Abglitt zum Schwa war ebenso bedeutend geringer als die im Abglitt zum Vokal /a/ (β = -9.738, SE = 4.419, t = -2.204). Die Interaktion von Sprechergruppe und Vokal hat den Model-Fit nicht verbessert ($\chi^2(2)=0.269$, p = 0.874).

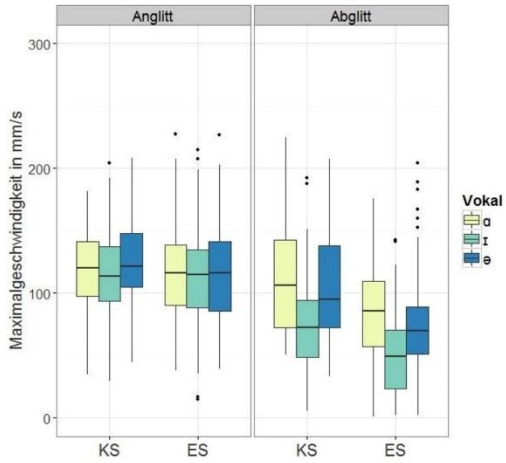


Abbildung 4: Maximalgeschwindigkeit von Anglitt und Abglitt in Abhängigkeit von Sprechergruppe (Erwachsene Stotterer/ES im Vergleich zu den Kontrollsprechern/KS) und Vokal (/a/, /i/, /ə/)

4. Diskussion

Unsere Ergebnisse zeigen, dass selbst die akustisch-perzeptuell flüssige Sprache von Stotterern sich in der Kinematik von der Sprache der Kontrollsprecher unterscheidet, was darauf hindeuten könnte, dass die akustisch-perzeptiven Stottereignisse lediglich die Spitze des Eisbergs darstellen. Die Bestimmung von Stottereignissen sollte daher nicht nur in der akustischen Charakterisierung bestehen, sondern kinematische Faktoren mit einbeziehen. Diese könnten eventuell unser Verständnis des Stotterns erweitern, indem sie zusätzliche Symptome (zu den drei vorherrschenden Hauptsymptomen) aufzeigen.

In der Lokalisierung einer möglichen „Fault-Line", konnten wir Anhaltspunkte finden, die die Hypothese Wingate's unterstützen. Wingate war der Annahme, dass nicht der initiale Konsonant selbst die Unflüssigkeiten im Sprechfluss auslöst. Stattdessen schlug er vor, dass Stottereignisse daher rühren, dass der folgende Vokal nicht problemlos integriert wird. Diese Annahme wird von unseren Daten unterstützt: Während die Sprechergruppen sich im Anglitt nicht wesentlich unterschieden, konnten wir für den Abglitt (Übergang vom konsonantischen Verschluss hin zum Vokal) sowohl für die Dauer, als auch für die erreichte Maximalgeschwindigkeit einen statistischen Effekt beobachten, was die Bedeutung der artikulatorischen Untersuchung unterstreicht.

Die Methode, die wir präsentiert haben, verbindet Aspekte der traditionell statischen Ultraschall-Datenanalyse mit denen von dynamischen Fleshpoint-Verfahren wie zum Beispiel der elektromagnetischen Artikulographie. Als Ausgangspunkt bietet das Ultraschallbild umfangreiche Informationen über die Zungenform im zweidimensionalen Raum. Die Aneinanderreihung mehrerer Ultraschallbilder bei hoher Bildrate bietet zudem eine zusätzliche temporale Dimension, die Einblicke in die Koordination der Artikulation bietet. Der Meßvektor wird an der Kinematik der Zunge selbst und somit an einem den Daten internen sprecherunabhängigen Referenzpunkt orientiert [6]. Dies berücksichtigt die artikulatorische und anatomische Heterogenität des einzelnen Sprechers, was eine relativ objektive Quantifizierung der kinematischen Kennwerte und somit Vergleiche von Sprechergruppen erlaubt.

Die Tatsache, dass wir mit dieser Methode selbst feine akustisch nicht erfassbare Unterschiede in der scheinbar flüssigen Sprache beider Gruppen festhalten und messen konnten, wird als Bestätigung dieser Methode interpretiert. Zudem unterscheiden sich die Messungen beider Sprechergruppen erwartungsgemäß in Abhängigkeit der verschiedenen Silbenkerne: Während im Anglitt keine signifikanten Unterschiede messbar waren, werden im Abglitt Effekte von Koartikulation ersichtlich, wobei sich Verläufe zum hohen Vokal /i/ durch eine signifikant geringere Dauer als auch geringere Maximalgeschwindigkeit hervorhoben. Letzteres bestätigt den Anspruch der Präzision dieser Methode.

5. Dank

Die Autoren danken Prof. Alan Wrench und Steve Cowen für deren stetige Unterstützung.

6. Bibliographie

[1] Wingate, M. E. (1988). The structure of stuttering: A psycholinguistic analysis. Springer Verlag, New York, NY.

[2] Alpermann, A. (2010) Redeflussstörung: Stottern–eine psychische Störung? Sprache Stimme Gehör, 34(2), 56

[3] Kaufmann, S. (2006) Idiopathisches Stotterin – Diskussion vor dem Hintergrund eines psycholinguistischen Modells der Sprachproduktion. München, GRIN Verlag.

[4] Tasko, S. M., & Westbury, J. R. (2002). Defining and measuring speech movement events. Journal of Speech, Language, and Hearing Research, 45(1), 127-142.

[5] Articulate Instruments Ltd (2012). Articulate Assistant Advanced User Guide: Version 2.14. Edinburgh, UK: Articulate Instruments Ltd.

[6] Iskarous, K. (2005). Patterns of tongue movement. Journal of Phonetics, 33(4), 363-381.

7.4 Appendix D: Heyde, Scobbie, Cleland & Roxburgh (2016)

UltraPhonix: Das Erlernen von Artikulatorischen Gesten mit Ultraschall-Biofeedback

Cornelia J Heyde [1], Joanne Cleland [2], James M Scobbie [1], Zoe Roxburgh[1]

[1] Clinical Audiology, Speech and Language (CASL) Research Centre, Queen Margaret University, Edinburgh, UK
[2] School of Philosophy, Psychology and Language Sciences (PPLS), Strathclyde University, Glasgow, UK

Queen Margaret University – Clinical Audiology, Speech and Language Research Centre

University of Strathclyde Glasgow

Hintergrundinformation

- zunehmende Popularität von Ultraschall als Biofeedback-Tool in der Therapie von Spracherwerbsstörungen (SES) [1, 2, 3]
- Visualisierung der Zungenbewegung in Realzeit als Teil der Sprachtherapie
- Überarbeitung des Motorprogrammes mit dem Ziel der Verbesserten Produktion des jeweiligen Ziellautes

Methode

- 20 Kinder (6.0 – 15.0) mit diversen SES, insbesondere mit Beeinträchtigung von lingualen Konsonanten oder Vokalen
- Interventionsstudie mit zehn Therapie-einheiten mit Ultraschall-Biofeedback (Abb. 1) [4]
- multiple Bewertungseinheiten (vor, während, direkt nach und drei Monate nach der Intervention; Tabelle 2)
- Produktion des Ziellautes in motorisch zunehmend komplexeren Kontexten (Tabelle 1) [1, 5]
- Wechsel auf motorisch komplexeres Niveau bei korrekter Produktion von 80% des Ziellautes

Abb. 1 – Setup der Therapiesitzung mit visuellem Ultraschall-Biofeedback

Tabelle 1
Motorisch komplexer werdende Niveaus

Level 0	CV oder VC erleichternde Vokale
Level 1	CV und VC
Level 2	CVC mit dem Ziellaut an Wortanfang und Wortende
Level 3	mehrsilbige Wörter
Level 4	Satzwiederholung mit Ziellaut an Wortanfang und Wortende
Level 5	Cloze Test (Vervollständigung von Sätzen)
Level 6	Cluster
Level 7	komplexe Sätze wiederholen und ausdenken

Tabelle 2
Inhalte der Therapie- u. Bewertungseinheiten

Baseline (*BL*) 1 – 3	Woche 1 - 3
• **Diagnostic Evaluation of Articulation and Phonology (DEAP)** Test: Bildbenennungstest mit CV-Kombinationen an Wortanfang und Wortende; Bestimmung des Ziellautes	
• **Unbehandelte Liste**: etwa 100 Wörter mit steigender Komplexität (Tabelle 1) und Sätze mit Ziellaut in verschiedenen Wortpositionen (Anfang, Mitte und Ende) und Vokalkontexten	
Therapiephase 1	**Woche 4 - 8**
• 5 Einheiten Ultraschall-Biofeedback-Therapie	
Midline (*Mid*)	**Woche 9**
• DEAP und unbehandelte Liste	
Therapiephase 2	**Woche 10 – 14**
• 5 Einheiten Ultraschall-Biofeedback-Therapie	
Postline (*Post*)	**Woche 15**
• DEAP und unbehandelte Liste	
Maintenance (*Maint*)	**Woche 30**
• DEAP und unbehandelte Liste	

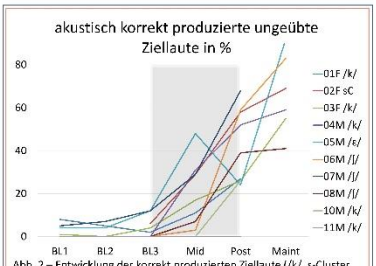


Abb. 2 – Entwicklung der korrekt produzierten Ziellaute (/k/, s-Cluster, /ɛ/, /ʃ/) von zehn Kindern (01F – 11M) mit diversen SES

Quantitative Analyse

- randomisierte Perzeptionsstudie mit zwei phonetisch trainierten Hörern
- akustische Bewertung des Ziellautes auf Korrektheit (J/N) (Abb. 2)

Beobachtungen

- genereller Trend zur korrekteren Produktion des Ziellautes (Abb. 2)
- stärkere Verbesserung nach Abschluss der zweiten Therapiephase (Post)
- messbar anhaltender Effekt (Maint)

Qualitative Analyse

- Splining, i.e., Einzeichnen der Zungenoberfläche in das Ultraschallbild (Abb. 3) relativ zum akustischen Signal (Burst bei /k/ und s-Clustern, Mittelpunkt bei /ʃ/ und /ɛ/)
- Überlagerung aller Splines eines Lautes und Berechnung des Mittelwertes, sowie der Standardabweichung der Zungenkonfiguration je Bewertungseinheit (Abb. 4)

Beobachtungen

- Veränderung der Zungenkonfiguration spiegelt akustische Verbesserung (Abb. 2) wider
- verschiedene Muster im Umbau der Artikulation: stetig bis Maintenance (03F – Abb. 4 links), beginnend mit Postline (05M – Abb. 4 mitte), abgeschlossen bis Postline (06M – Abb. 4 rechts)

Abb. 3 – Ultraschallbild mit mid-saggitaler Zungenkontur

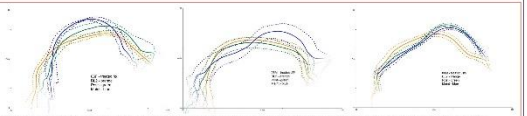


Abb. 4 – Mittelwerte (durchgehende Linie) und Standardabweichung (gestrichelte Linie) für die drei Bewertungseinheiten vor (BL2 - orange), direkt nach (Postline - grün) und drei Monate nach (Maintenance - blau) Abschluss der Intervention für drei Probanden (03F, 05M, 06M)

Diskussion

- deutlich erkennbarer positiver Effekt von Ultraschall-Biofeedback in der Therapie von lingualen Sprachentwicklungsstörungen bei Kindern mit diversen SES
- Ultraschall in der Therapie: Relativ intuitive Darstellung der Zunge als zusätzlicher Kanal um neue Artikulationen zu erlernen bzw. bestehende Artikulationsmuster zu überschreiben
- Ultraschall in Diagnostik und Analyse: Mittelwerte der Zungenkonfiguration lassen auf bestehende Muster in der Artikulation, sowie Veränderungen im Artikulationsmuster schließen; Standardabweichungen auf die Präzision der Ausführung einer Artikulation

Referenzen:
[1] Cleland, J., Scobbie, J.M. & Wrench, A., (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. Clinical Linguistics and Phonetics. Pp. 1-23.
[2] Gibbon, F. E. (1999). Undifferentiated lingual gestures in children with articulation/phonological disorders. Journal of Speech, Language, and Hearing Research, 42(2), 382-397.
[3] Heng, Q., McCabe, P., Clarke, J., & Preston, J. L. (2016). Using ultrasound visual feedback to remediate velar fronting in preschool children: A pilot study. Clinical linguistics & phonetics, 1-16.
[4] Articulate Instruments Ltd (2012). Articulate Assistant Advanced User Guide: Version 2.14. Edinburgh, UK: Articulate.
[5] Preston, J.L., McCabe, P., Rivera-Campos, A., Whittle, J.L., Landry, E., Maas, E. (2014) Ultrasound visual feedback treatment and practice variability for residual speech sound errors. Journal of Speech, Language, and Hearing Research 57(6), 2102-15.

Kontakt:
Cornelia J Heyde
cheyde@qmu.ac.uk

UltraPhonix: Das Erlernen von Artikulatorischen Gesten mit Ultraschall-Biofeedback

Cornelia Heyde[1], Joanne Cleland[2], James Scobbie[1], Zoe Roxburgh[1]

[1] Queen Margaret Universität, Edinburgh, UK
[2] Strathclyde Universität, Glasgow, UK

1 Zusammenfassung

In einer Interventionsstudie mit 20 Kindern wird die Effektivität von visuellem Ultraschall-Biofeedback in der Therapie verschiedener Sprechstörungen untersucht. Die Studie umfasst mehrere standardisierte Tests, um die Sprach-, und Sprechfähigkeiten der Kinder vor, während und nach der sprachtherapeutischen Intervention (in multiplen *Baselines*) zu erheben. Alle Teilnehmer waren zum Zeitpunkt der ersten Aufnahme zwischen 6.0 und 15.0 Jahren alt und wiesen Sprachentwicklungsstörungen mit einer persistierenden Symptomatik auf der phonologisch-phonetischen Ebene auf, die sich in einer Vielzahl von Konsonant- und Vokalfehlern widerspiegelten. Die Produktion der Ziellaute wurde an Übungsitems trainiert und deren Genauigkeit anhand von ungeübten Wörtern (Kontrollitems) beobachtet und ausgewertet. Klinisch bedeutende Verbesserungen direkt im Anschluss und drei Monate nach Beendigung der Ultraschall-Biofeedback Therapie werden im Vergleich zur den Aufnahmen vor Therapiebeginn vorgestellt.

2 Hintergrundinformation

Ultraschall als Biofeedback-Tool bietet eine visuelle Unterstützung in der Sprachtherapie indem die Zunge in Realzeit abgebildet wird. Besonders in der Therapie von (insbesondere persistierenden) Sprachentwicklungsstörung gewinnt Ultraschall zunehmend an Popularität. Medizinische Standardultraschallgeräte werden verwendet um die Zungenbewegungen (mit sagittalem/seitlichem oder koronalem/fron-

talem anatomischem Schnittbild; cf. Abb. 1 und 2) in Realzeit aufzunehmen und darzustellen. Dies bietet dem Patienten (zusätzlich zu der propriozeptiven Wahrnehmung) visuelle Information über die Bewegung der Zunge, die hilft, die artikulatorische Koordinierung besser zu reflektieren und zu verstehen (Preston, Brick & Landi, 2013) und somit die zuvor fehlerhaft ausgeübte Artikulation von Lauten nun mit überarbeiteten, präziseren Motoprogrammen zu produzieren.

Wissenschaftlich ist der Nutzen von visueller Ultraschall-Biofeedback Therapie bisher noch nicht ausreichend erwiesen - mit über 20 Einzelfall- oder kleineren Gruppenstudien jedoch recht vielversprechend. In vielen Fällen wird Ultraschall verwendet um die verzögerte oder gestörte Produktion des Lautes /r/ zu untersuchen (cf. Bacsfalvi, 2010, McAllister, Hitchcock & Swartz, 2014, Preston, McCabe, Rivera-Campos, Whittle, Landry & Maas, 2015). In einigen neueren Studien wurde die Bandbreite der untersuchten Ziellaute erweitert. So wurde Ultraschall zum Beispiel für die Untersuchung von velaren und alveolaren Plosiven oder Sibilanten verwendet (Cleland, Scobbie & Wrench, 2015, Heng, McCabe, Clarke & Preston, 2016).

Die vorliegende Studie untersucht den Nutzen von Ultraschall in der visuellen Biofeedback-Therapie von heterogenen persistierenden Sprachentwicklungsstörungen bei Kindern. Die Annahme, die dieser Studie zugrunde liegt ist, dass Ultraschall-Biofeedback die Sprachproduktion bei vorliegender phonetischer Sprech- (z.B. Sigmatismus, Kappazismus oder Vokalstörungen), wie auch phonologischer Sprachstörung (z.B. Auslassung oder Reduktion, Alveolarisierung, Velarisierung) positiv beeinflusst. Zwei Stufen der Sprachlautentwicklung werden unterschieden: (1) der Erwerb, wenn ein Kind die Artikulation des Ziellautes erstmals erfolgreich realisiert und (2) die Generalisierung, wenn die neu erworbene artikulatorische Geste in unbekanntem Kontext produziert wird (Schmidt & Lee, 1999). Für die akustische, wie auch die artikulatorische (Ultraschall-) Analyse der Daten wird die Produktion von Übungsitems und Kontrollitems ausgewertet.

3 Methode

Zwanzig Kinder im Alter von 6.0 bis 15.0 Jahren mit persistierenden Sprachentwicklungsstörungen (ohne Anzeichen von anatomisch-strukturellen Anomalitäten), die die Produktion von Vokalen und/oder lingualen Konsonanten beeinflussen, haben an der Studie teilgenommen. Die Datenerfassung ist noch nicht abgeschlossen. Zum derzeitigen Zeitpunkt haben sechs Probanden die Therapie abgeschlossen, fünf Probanden wurden von der Therapie ausgeschlossen und weitere neun Probanden befinden sich in der Endphase und werden bis voraussichtlich Februar 2017 die Therapie abschließen.

Die Studie ist eine Interventionsstudie mit multiplen Baseline-Erhebungen (cf. Tabelle 1) um den Entwicklungsstand vor, während und nach der Therapie zuverlässig feststellen zu können. Wortlisten (mit ungeübten Wörtern) wurden auf die Sprachentwicklungsstörung jedes Probanden individuell abgestimmt. Probanden wurden von der Teilnahme an der Studie ausgeschlossen, wenn sie bei den Aufnahmen vor Therapiebeginn (Baseline 1, 2 und 3 im Durchschnitt) 20% oder mehr der Ziellaute in der unbehandelten Wortliste korrekt produzieren konnten. Parallel zu den Audiodaten wurden Ultraschalldaten der Zunge aufgenommen.

Während der Therapiephase wurde die Ultraschallsonde unter dem Kinn des Probanden mit einem Headset stabilisiert und das Ultraschallsignal wurde auf einen Bildschirm projiziert, wo es in Realzeit die Bewegungen der Zungenoberfläche (weiße Linie auf ansonsten schwarzem Hintergrund) abbildete. Eine Referenzlinie für den harten Gaumen wurde auf den Bildschirm gezeichnet um einen visuellen anatomischen Kontext zu gewinnen, der die Interpretation der Zungenbewegung für Proband, wie auch Therapeut erleichterte. Diese Referenzlinie wurde benötigt um selbst subtile artikulatorische Unterschiede, wie die von hohen und tiefen Vokalen oder von uvularen und velaren Konsonanten, sichtbar zu machen. Die Ultraschalldaten bieten dem Therapeuten Informationen über die Koordinierung der Zunge, was die Diagnose von Sprachfehlern erleichtert.

Tabelle 1
Zeitverlauf und Inhalte der Therapie- und Bewertungseinheiten

Baseline 1 - 3	**Woche 1 - 3**
• ***Diagnostic Evaluation of Articulation and Phonology*** *(DEAP) Test - ein Bildbenennungstest mit 50 Wörtern, die alle Konsonanten in am Wortanfang und Wortende mit allen Vokalen kombinieren: zur Bestimmung des Ziellautes für die sprachtherapeutische Intervention* • *Unbehandelte Liste: etwa 100 Wörter mit steigender Komplexität (CV, VC, CVC, CCVC bis zu mehrsilbigen Wörtern) und 10 Sätze, die den Ziellaut in verschiedenen Wortpositionen (am Anfang, Mitte und Ende des Wortes) und in verschiedenen Vokalkontexten beinhalten; Wörter wurden gelesen oder nachgesprochen* • ***Intelligibility in Context Scale*** (ICS): Fragebogen, der die Verständlichkeit der Kinder mit Personen im Familienkreis, wie auch mit unbekannten Personen wiederspiegelt	
Therapiephase 1	**Woche 4 - 8**
• 5 Einheiten visuelle Ultraschall-Biofeedback Therapie	
Midline	**Woche 9**
• DEAP (phonologischer Teil), unbehandelte Liste, ICS (siehe Baseline 1-3)	
Therapiephase 2	**Woche 10 – 14**
• 5 Einheiten visuelle Ultraschall-Biofeedback Therapie	
Postline	**Woche 15**
• DEAP (phonologischer Teil), unbehandelte Liste, ICS (siehe Baseline 1-3)	
Maintenance	**Woche 30**
• DEAP (phonologischer Teil), unbehandelte Liste, ICS (siehe Baseline 1-3)	

Während der Therapie (nicht jedoch während der Baseline-Aufnahmen) konnte der Proband die Oberflächenkontur seiner Zunge auf dem Ultraschallbildschirm sehen. Bewegungen der Zunge wurden in Realzeit mit hoher zeitlicher Auflösung (etwa 120 Bilder pro Sekunde) auf dem Bildschirm widergegeben. Diese visuelle Information über die Artikulation des Probanden wurde von dem Sprachtherapeuten genutzt um Fehlartikulationen zu identifizieren und um diese mit dem Probanden gemeinsam zu korrigieren. Hierbei sollte die visuelle Darstellung dem Probanden ermöglichen, seine Artikulation unmittelbar zu modifizieren. Die Therapiemethode basierte auf den Prinzipien des motorischen Lernens (cf. Cleland et al., 2015 und Preston et al., 2013) mit Übungsitems, die motorisch zunehmend anspruchsvoller wurden (cf. Tabelle 2).

Abbildung 1. Ultraschallbild mit sagittaler Perspektive

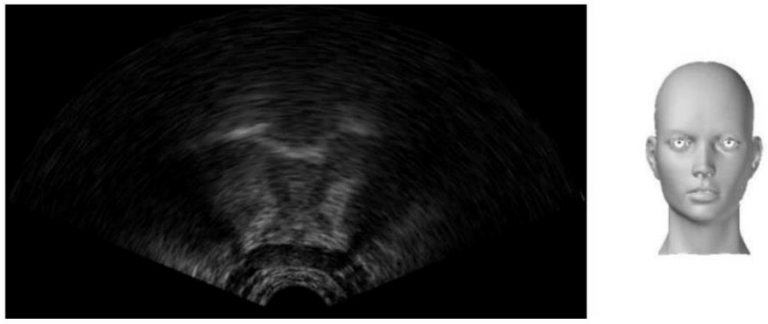


Abbildung 2. Ultraschallbild mit koronaler Perspektive

Tabelle 2

Niveaus mit zunehmend komplexer werdender Motorik

Level 0	CV oder VC erleichternde Vokale
Level 1	CV und VC
Level 2	CVC mit dem Ziellaut an Wortanfang und Wortende
Level 3	mehrsilbige Wörter
Level 4	Satzwiederholung mit dem Ziellaut an Wortanfang und Wortende
Level 5	Cloze Test (Vervollständigung von Sätzen)
Level 6	Cluster
Level 7	komplexe Sätze wiederholen und ausdenken

Die Kinder begannen mit einem einfacheren Niveau und bei korrekter Produktion von 80% des individuell abgestimmten Ziellautes in den Übungsitems, wurde auf ein motorisch anspruchsvolleres Niveau gewechselt (z.B. von einsilbigen auf zweisilbige Wörter). Dieser schrittweise Ansatz erlaubte es uns zu beobachten, wie lange es im Durchschnitt dauerte, a) eine neue artikulatorische Geste zu erlernen und b) diese zu generalisieren und auch in ungeübten Wörtern (Kontrollitems) zu bilden. Insgesamt erhielt jedes Kind zehn Therapieeinheiten mit visuellem Ultraschall-Biofeedback mit drei vorangestellten Baseline-Aufnahmen, einer Zwischenstandsaufnahme (Midline) und zwei der Therapie folgenden Baseline-Aufnahmen, wovon eine direkt im Anschluss an die Therapie erfolgte (Postline) und die zweite drei Monate nach Beendung der Ultraschall-Biofeedback Therapie (Maintenance), die der Nachhaltigkeitsreflektion (cf. Tabelle 1) dient.

4 Analyse der Daten

Bevor die Daten analysiert werden konnten, wurden sie von einem erfahrenen Sprachtherapeuten eng phonetisch transkribiert. Die Transkription wurde blind durchgeführt, was bedeutet, dass der Transkriptor keine Informationen darüber erhielt, zu welchem Zeitpunkt (ob vor, während oder nach der Ultraschall-Biofeedback Therapie) die Daten erhoben wurden. Für die Ziellaute wurde der Prozentsatz der korrekt produzierten Laute berechnet (cf. Abb. 4). Für die Analyse der Ultraschalldaten wurden alle Wiederholungen eines Ziellautes annotiert – bei Plosiven die Öffnung des Verschlusses und bei Frikativlauten oder Sonoranten der akustische Mittelpunkt. Für jede Annotation wurde eine Linie (Spline), die der Zungenoberfläche nachempfunden ist, in das Ultraschallbild eingezeichnet. Für die Analyse wurden die verschiedenen Splines verglichen – sowohl innerhalb einer Therapieeinheit, als auch über mehrere Therapieeinheiten hinweg (cf. Abb. 3).

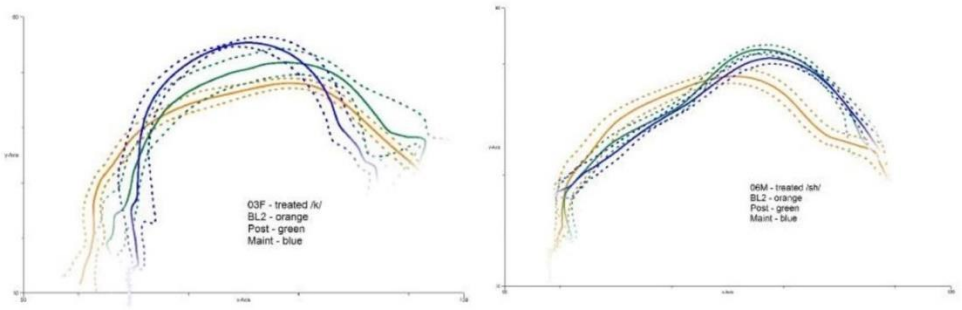


Abbildung 3. Splines der Zungenoberfläche (Zungenspitze rechts, Zungenwurzel links) des Ziellautes von der Baseline (orange) über die Postline (grün) bis hin zur Maintenance (blau) für die beiden Probanden 03F (Ziellaut /k/) und 06M (Ziellaut /ʃ/).

So wurde anhand von Minimalpaaren (zum Beispiel /top/, /cop/ bei Alveolarisierung) untersucht, (1) ob sich die artikulatorischen Gesten für verschiedene Ziellaute (in diesem Fall /t/ und /k/) innerhalb einer Einheit unterschieden und (2) ob sich die artikulatorischen Gesten nach der Therapie im Vergleich zu vor der visuellen Ultraschall-

Biofeedback Therapie verbesserten. Die Abbildungen der Zungenkontur wurden so auf einen eventuellen verdeckten Kontrast, i.e., Covert Contrast (differenzierte Artikulation bei gleichem akustischem Lautbild) oder nicht-differenzierte Zungenkonturen untersucht (Gibbon, 1999).

5 Ergebnisse

Die akustischen Ergebnisse der Probanden, die die Therapie bereits abgeschlossen haben, zeigen einen relativ schnellen Erwerb von neuen Artikulationen innerhalb der ersten Therapieeinheit mit Ultraschall-Biofeedback. Die Generalisierung dieser neu erlernten artikulatorischen Gesten auch auf ungeübte Kontrollwörter tritt hingegen etwas verzögert ein.

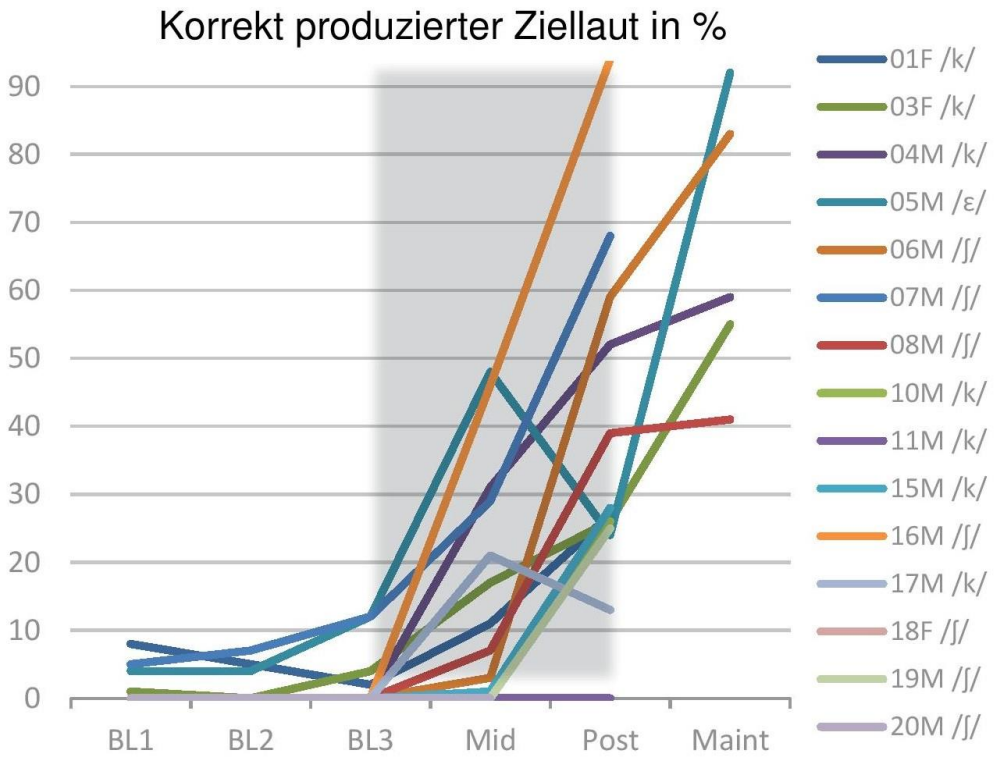


Abbildung 4. Entwicklung der korrekt produzierten Ziellaute von vor der Intervention bis zur Postline direkt nach Abschluss der Therapie

Von den Kindern, die die Postline (Aufnahme direkt nach der letzten Therapieeinheit) abgeschlossen haben, konnten alle eine deutliche Verbesserung in der Anzahl der korrekt produzierten Ziellaute (Konsonanten oder Vokale) zeigen. Im Durchschnitt konnte eine perzeptuelle Verbesserung von 37% im Vergleich zu den Aufnahmen vor Therapiebeginn beobachtet werden – mit weiterer Besserung in der Maintenance (Aufnahme drei Monate nach Abschluss der Therapie). Artikulatorisch deuten die bereits ausgewerteten Ultraschalldaten nicht auf Covert Contrasts (akustisch nicht wahrnehmbare Kontraste) in der Produktion der Minimalpaare hin. Stattdessen wurden nicht-differenzierte Zungenkonturen bei den Kindern mit Alveolarisierung und einige abnorme Zungenkonturen beobachtet. Die akustisch-perzeptiven Verbesserungen in der Postline werden in signifikanten Veränderungen der Zungenkonfiguration widergespiegelt.

6 Diskussion

Mit unserer Studie konnten wir einen vorläufigen Beleg für die Wirksamkeit von visuellem Ultraschall-Biofeedback in der Therapie von Sprechlauten bei Kindern erbringen. Zu Beginn der Studie zeigten die Probanden eine Vielzahl an persistierenden Sprachentwicklungsstörungen. Mithilfe eines Headsets wurde die Position der Ultraschallsonde stabilisiert, was dazu führte, dass sich die Zunge in Bezug auf die palatal eingezeichnete Referenzlinie wesentlich stabiler verhielt. Weiterhin ermöglichte die Stabilisierung der Sonde den Vergleich von Zungenkonfigurationen zwischen mehreren Aufnahmen vor und nach der Ultraschall-Biofeedback Therapie. Die Analyse dieser Zungenkonfigurationen lässt eine Vielzahl an abnormen artikulatorischen Gesten, einschließlich der undifferenzierten Gesten, erkennen, was (zumindest für die Probanden dieser Studie) auf eine motorische Ursache der persistierenden Sprach-Laut-Störungen hinweist und der Diagnose „phonologische Störung“, mit der einige der Kinder an uns überwiesen wurden, entgegensteht.

7 Literatur

Bacsfalvi, P. (2010). Attaining the lingual components of /r/ with ultrasound for three adolescents with cochlear implants. *Journal of Speech-Language Pathology and Audiology*, 34(3), 206-217.

Cleland, J., Scobbie, J.M. & Wrench, A., (2015). Using ultrasound visual biofeedback to treat persistent primary speech sound disorders. *Clinical Linguistics and Phonetics*. 1-23.

Gibbon, F. E. (1999). Undifferentiated lingual gestures in children with articulation/phonological disorders. *Journal of Speech, Language, and Hearing Research*, 42(2), 382-397.

Heng, Q., McCabe, P., Clarke, J. & Preston, J. L. (2016). Using ultrasound visual feedback to remediate velar fronting in preschool children: A pilot study. *Clinical linguistics & Phonetics*, 1-16.

McAllister Byun, T. M., Hitchcock, E. R. & Swartz, M. T. (2014). Retroflex versus bunched in treatment for rhotic misarticulation: Evidence from ultrasound biofeedback intervention. *Journal of Speech, Language, and Hearing Research*, 57(6), 2116-2130.

Preston, J. L., Brick, N. & Landi, N. (2013). Ultrasound biofeedback treatment for persisting childhood apraxia of speech. *American Journal of Speech-Language Pathology*, 22(4), 627-643.

Preston, J.L., McCabe, P., Rivera-Campos, A., Whittle, J.L., Landry, E. & Maas, E. (2014) Ultrasound visual feedback treatment and practice variability for residual speech sound errors. *Journal of Speech, Language, and Hearing Research,* 57(6), 2102-15.

Schmidt, R. A. & Lee, T. D. (1999). Motor Control and Learning (3. Ed.). Champaign, IL: *Human Kinetics*.

7.6 Appendix F: Information Sheet for PWS

Queen Margaret University

EDINBURGH

Division of Speech and Hearing Sciences

Information for Ultrasound Tongue Imaging Subjects Who Stammer

We are currently carrying out detailed analysis of the speech of people who stammer. This involves making UTI (Ultrasound Tongue Imaging), sound and video recordings in the Speech Production Research Laboratory at Queen Margaret University, Musselburgh. Recordings will be done in two one-hour sessions with a longer break between sessions. This gives enough time for us to assess the severity of your stammer and to set you up with the ultrasound machine and record your speech. You will be asked to read syllables and sentences off a computer screen in a relaxed and natural way.

One way to get an idea of what will happen is to view the equipment in the lab. You will be shown the equipment before the actual recording and encouraged to ask questions and discuss what will happen. This sheet is a summary.

You are free to call a halt at any stage during data collection

There is no obligation on you to take part. You are free to withdraw before or during the recording. It is important for us, and for you, that you feel relaxed and as comfortable as possible during the recording.

What do we need from you?

You must be a native speaker of English of 18 years of age or older. You should have a stammer, which occurred when you were still a child (typically by the age of 8 years of age). You should have no other speech, language, hearing or visual impairment that could influence the study. You should be in general good health and be able to sit upright in a chair for an hour without danger of back ache, loss of circulation to the legs or other unpleasant effect. You should be able to tolerate a little discomfort because we need you to wear a stabilisation helmet during the recording.

What is the experiment for?

We want to find out how the tongue moves about in the mouth and how it changes shape during fluent and disfluent speech of people who stammer. We also want to discover how certain consonants and vowels might influence the way your tongue moves. The ultrasound probe will capture these movements of your tongue. In addition, a camera attached to the stabilisation helmet will capture video data of your lips, which is useful to obtain information on when you start to speak.

What will happen during the session?

Before the Ultrasound recording we will employ a standardised speech assessment tool to assess the severity of the stammer. This should last no more than 15 minutes. We will then go on to the actual data collection using ultrasound tongue imaging. During the ultrasound recording you will sit in an adjustable office-type chair inside our small recording studio. A technician will put on the stabilisation helmet with the attached lip camera. The helmet is adjusted to fit, and the ultrasound probe will be positioned.

Figure 1 (left) Ultrasound machine and transducer (handheld probe), (right) the ultrasound probe and headset

When comfortable, we ask you to read a number of syllables (approximately 200) with a short pause in between each syllable as the data is stored on the computer. You will be instructed to read the syllables in either regular or whispered mood. Another condition requires you to read phrases off the screen. Two separate sessions will allow you to take a longer break. Each session consists of about 12 blocks. After each block you are free to have a longer pause. You are also fairly free to move in pauses as the helmet will compensate for movements. The researcher will be in the neighbouring room and available to answer questions at all times. A technician will also be available during the recording to help in case of technical difficulty. The recording may become fairly dull and repetitive, so we will check that you have frequent pauses. The room can get warm, so light clothes are preferable. When the session is finished, you're helped to get out of the stabilisation helmet. Slight marks on your face might be visible for a couple of minutes after the recording. A little stretching and tensing helps after sitting still for so long. You'll be invited to view your data on the computer.

What do we do with the data?

These recordings of your speech will be kept securely and used solely for research and teaching purposes. We hope to publish the results of the research in the future in books, journals and online, but complete confidentiality will be respected at all times. Your name will not be used, your face will be obscured in any photos, and you will not be identified as the speaker in any way other than in your consent form, which we will keep securely.

If you are willing to participate in this project, please sign the attached consent form and return it to the researcher. Please keep this information sheet. You are encouraged to get in touch afterwards if you have further questions:

Cornelia J Heyde Email: cheyde@qmu.ac.uk
Tel.: 07510552477

7.7 Appendix G: Information Sheet for PNS

Queen Margaret University

EDINBURGH

Division of Speech and Hearing Sciences

Information for Ultrasound Tongue Imaging Subjects

We are currently carrying out detailed analysis of the speech. This involves making UTI (Ultrasound Tongue Imaging), sound and video recordings in the Speech Production Research Laboratory at Queen Margaret University, Musselburgh. Recordings will be done in two 30-minute sessions with a longer break between sessions. You will be asked to read syllables and sentences off a computer screen in a relaxed and natural way.

One way to get an idea of what will happen is to view the equipment in the lab. You will be shown the equipment before the actual recording, and encouraged to ask questions and discuss what will happen. This sheet is a summary.

You are free to call a halt at any stage during data collection

There is no obligation on you to take part. You are free to withdraw before or during the recording. It is important for us, and for you, that you feel relaxed and as comfortable as possible during the recording.

What do we need from you?

You must be a native speaker of Scottish or Irish English of 18 years of age or older. You should have no speech, language, hearing or visual impairment that could influence the study. You should be in general good health and be able to sit upright in a chair for an hour without danger of back ache, loss of circulation to the legs or other unpleasant effect. You should be able to tolerate a little discomfort because we need you to wear a stabilisation helmet during the recording.

What is the experiment for?

We want to find out how the tongue moves about in the mouth and how it changes shape during speech. We want to discover how certain consonants and vowels might influence the way your tongue moves. The ultrasound probe will capture these movements of your tongue. In addition, a camera attached to the stabilisation helmet will capture video data of your lips, which is useful to obtain information on when you start to speak.

What will happen during the session?

During the ultrasound recording you will sit in an adjustable office-type chair inside our small recording studio. A technician will put on the stabilisation helmet with the

attached lip camera. The helmet is adjusted to fit, and the ultrasound probe will be positioned.

Figure 1 (left) Ultrasound machine and transducer (handheld probe), (right) the ultrasound probe and headset

When comfortable, we ask you to read a number of syllables (approximately 350) with a short pause in between each syllable as the data is stored on the computer. You will be instructed to read the syllables in either regular or whispered mode. Another condition requires you to read phrases off the screen. Two separate sessions will allow you a longer break to rest. Each session will last no longer than 35 minutes. Separate blocks allow you to have pauses. You are also fairly free to move in pauses as the helmet will compensate for movements. The researcher will be in the neighbouring room and available to answer questions at all times. A technician will also be available during the recording to help in case of technical difficulty. The recording may become fairly dull and repetitive, so we will check that you have frequent pauses. The room can get warm, so light clothes are preferable. When the session is finished, you're helped to get out of the stabilisation helmet. Slight marks on your face might be visible for a couple of minutes after the recording. A little stretching and tensing helps after sitting still for so long. You'll be invited to view your data on the computer.

What do we do with the data?

These recordings of your speech will be kept securely, and used solely for research and teaching purposes. We hope to publish the results of the research in the future in books, journals and online, but complete confidentiality will be respected at all times. Your name will not be used, your face will be obscured in any photos, and you will not be identified as the speaker in any way other than in your consent form, which we will keep securely.

If you are willing to participate in this project, please sign the attached consent form and return it to the researcher. Please keep this information sheet. You are encouraged to get in touch afterwards if you have further questions:

Cornelia J Heyde Email: cheyde@qmu.ac.uk
Tel.: 07510552477

7.8 Appendix H: Consent Form

Queen Margaret University

EDINBURGH

Division of Speech and Hearing Sciences

Consent Form for Ultrasound Tongue Imaging Subjects

I have read and understood the consent form and the information sheets, and seen the equipment, which together give information about the project, the experiment, and the laboratory. I had the opportunity to ask questions about them.
I understand I am under no obligation to take part in this study and a decision not to participate will not be a problem. I understand that I have the right to withdraw from this study at any stage before or during data collection, without giving any reason. I understand that this is non-therapeutic research from which I cannot expect to derive any benefit.

I fully agree to the necessary conditions:

☑ I agree to participate in this experiment.
☑ I agree that any audio and visual ultrasound, video and photographic data can be stored and used indefinitely but anonymously for analysis, research, academic conference presentations, and future applications for research funding, and that the anonymous results of the study can be disseminated freely to audiences and research users of all types.

Optionally, I agree that:

☐ anonymous recordings of my voice and visual images from ultrasound and video can be used in university teaching.
☐ anonymous recordings of my voice and visual images from ultrasound and video can be played to a public audience to advance understanding of science, through the internet, broadcast, laboratory and open days, science festivals and other public but non-professional talks and presentations.
☐ the anonymous raw ultrasound, video and audio data can be copied for analysis by other researchers outside QMU for their own academic research projects.

Name ..... Date of Birth ..... Sex.....
Address ....
Phone number.....
Signature ..... Date .....
Signature of Researcher .....
Further information is available from:

Cornelia J Heyde Email: cheyde@qmu.ac.uk
Tel.: 07510552477

7.9 Appendix I: Online Questionnaire

Name _____ Date _____

Questionnaire for people who stammer taking part in the ultrasound study

Please fill in the questionnaire below. The responses you submit to this questionnaire will be treated as strictly confidential. After filling in the questionnaire, please do not forget to press the 'Submit' button; otherwise we will not receive it.

1. Email
2. Landline telephone
3. Mobile telephone
4. Age
5. Gender *
   - male
   - female
6. Handedness *
   - right-handed
   - left-handed
7. What is your current occupation?
8. What is the highest level of educational qualification you have?
   (If your qualifications differ from those listed, please select the nearest match) *
   - GCSE
   - 'A' level
   - Diploma
   - First Degree
   - Post-graduate degree
9. Where did you grow up?
10. If you did not grow up in the UK how long (in years) have you been to the UK?
11. What is your first language?

12. Are you bilingual (or more)?
    - Yes
    - No
13. If English is not your first language at what age did you acquire it?
14. If English is not your first language how proficient would you say you are?

	1	2	3	4	5	
Poor	○	○	○	○	○	**Excellent**

15. How would you describe your English accent (e.g., Glaswegian, Australian)?
16. Approximately, how old (in years) were you when you started stammering?
17. How many months did your longest ever experience of remission last?
18. Has your stammer been diagnosed by a Speech and Language Therapist?
    - Yes
    - No
19. Have you ever had any other condition affecting your speech or communication (e.g., lisp, apraxia of speech)?
    - Yes
    - No
20. If yes, then please give very brief details (not more than 100 words).
21. Have any of your family (brothers, sisters, parents or children) ever had any condition affecting their speech or communication (e.g., stammer)?
    - Yes
    - No
22. If yes, then please give very brief details (not more than 100 words).
23. Have you ever had any therapy for stammering?
    - Yes
    - No
24. If yes, then please give very brief details (not more than 100 words).
25. Are there any particular words or situations you go out of your way to avoid?
    - Yes
    - No

26. If yes, then please give very brief details (not more than 100 words).

7.10 Appendix J: Passage used for SSI-IV Reading Task

The talk over salad and fromage was about ghosts. My English friend Christopher Neville informed me that two of them haunt his house in southern France, on the sunny terrace of which we were now having lunch. I don't normally believe in spirits, but it seemed wise to suspend disbelief for the moment, since I would soon be entering a region of sorcery and hidden Grails, where heretics once marched defiantly into the bonfires of bloodthirsty crusaders: the land of the Cathars. Christopher's ghosts were said to be knights from those medieval times. I don't know whether he began studying the Middle Ages because of the ghosts or whether the ghosts arrived one day because he had taken an unusually keen interest in the Cathars. I do know that his knowledge proved invaluable.

The Cathars, I had read, were a kind and gentle people. They were dualists (man is bad, the spirit good), they viewed the material world as corrupt, and they rejected certain important tenets of the powerful Catholic Church, including priests, the Trinity and the sacraments. The laying on of hands was thought to transform believers into the "Perfects" or Good Christians, who were from then on expected to abstain from sex and meat. The popularity of this gnostic faith threatened the reign of Pope Innocent III. In 1208, he sent Simon de Montfort on a crusade against the heretics. The crusade took its name from the town of Albi and was followed 25 years later by the Inquisition.

Note. Excerpt from "In Pursuit of the Perfect Ham in Little-Known Friuli," by Russ Parsons, *The Los Angeles Times*, March 19, 2000. Copyright 2000, Los Angeles Times. Reprinted with permission.